TitlePage

Elements of Statistics

Primary Author: Paul Flesher

Secondary Author: Jeffrey Sadler

Chief Editor: Jonathan Rehmert

Readability Editor: Lanee Young

Acknowledgements: We are very grateful for the work of the LibreText community for housing this project and facilitating the work of open educational resources. We began this project as a remixing project of David Lane's Introductory Statistics book which came from his original work over at onlinestatbook. As we began to develop our own voice and style, the project became less of a remix and more of a creation in our own voice. At each stage, however, David Lane's work served as inspiration and guide; despite having stark contrasts, we cannot separate our work from his (see the links throughout the book to see his original work and influence on our text); we appreciate his presentation and expertise. In most of the content, our conclusions align, but at times there were philosophical and mathematical differences. At these junctures, the disparities are made known and expounded upon to the degree we thought appropriate for an introductory text. We are thoroughly grateful for Lane's work; his team deserves a lot of credit. We have been using his sampling distribution simulations for quite some time. We must also acknowledge the people over at Hawkes Learning; we have been using their products in the classroom for years. Their influence has no doubt had major influence on the production of this book and our vision for the course.

Special thanks to Bill Weber for constructing the Excel guides and Jailynn Froese for creating all of the section Excel files.

Online homework for this text can be found at MyOpenMath.

Special acknowledgement to Jayme Goetz for compiling all of the online homework.

Please reach out to Paul Flesher at pmflesher@fhsu.edu with feedback or questions.

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

TABLE OF CONTENTS

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

Student Usage of the Book

Studying Mathematics and Statistics

To understand mathematical and statistical content proficiently, consistent dedication of two of our most precious and demanded resources, time and attention, are necessities. Both time and attention are in high demand and seem to be scarce commodities. We must be intentional in setting and keeping a schedule; that is our time commitment. Given today's society and technology, we must be even more intentional in our attention. Our thoughts, phones, computers, and family/friends consistently interrupt and with great frequency. We can mitigate the influence of the last three preemptively by selecting an appropriate location, setting boundaries, closing applications, and turning off notifications. Some of us might actually miss the constant stream of interruptions when they are gone (for our brief study sessions); we are not used to being in silence, thinking deeply, or attending for lengthy durations. Many of us will struggle; that is okay. We are not alone in this.

We cannot separate ourselves from our thoughts, and the concerns of the day often creep up when trying to study. We recommend a couple practices. There are many possibilities; find some that work for you. When you first sit down to study, take a minute or two to settle your mind, acknowledge the fact that your concerns are real but not pressing; they can wait until after your study session. During the study session, when you realize your mind is wandering or distracted, acknowledge that fact and then immediately reorient back to the material. We cannot hope to eliminate all mental distractions, but we can try to minimize the time that we are distracted.

Mathematical and statistical content takes attention and practice to understand. When we read, take notes, work examples, and attempt problems, we are working towards understanding and internalizing the ideas. We are not memorizing procedures or merely crossing tasks off a to-do list. Intention matters tremendously; remember why we are studying. Perhaps, we do not know why. If that is the case, begin pondering and researching. We address this in part early in the first chapter. For this semester, let us set the time and space for us to attend to course material so that we can understand the ideas, beauty, and applications proficiently.

Book: Reading, Notes, and Exercises

Both mathematics and statistics courses build throughout the semester. There is a logical progression; there is a story that builds continuously. Textbooks often chunk the story into disparate bites to the point that the story is lost. We have tried earnestly to relay the story as well as we can. If you do not understand something, do not move on as if it does not matter: identify that which does not make sense and ponder, reread, ask questions, etc. We strongly encourage you to take notes. Each lesson has learning objectives to help gauge what is important and bolds key terms to help you recognize their importance and for ease if you need to return for reference (ideally, your notes will work better for a quick reference). To be successful, we need both the big story and the details; it is important to attend to both. To help achieve this goal, we have created many text exercises.

? Text Exercise 1

When you see a box like this, it means that we have introduced some concept or idea that we think you are ready to think about and explore. Read the prompt and engage with it. We often ask for justification. Construct an answer in your notebook if you can. Compare your answer to our answer. Make modifications to your understanding. Go back and see if you can identify something that you missed that was important to the solution.

Answer

Do not read the prompt and immediately click to see the answer. Trying these exercises on your own is key to understanding and internalizing the material. If you cannot come up with a solution, that is not a problem (some of them are harder than others). Use the answer to help you. Try to identify the ideas in the answer and then go back to where they were presented in the text. Ask yourself: what could I have done to help draw these connections?

Learning statistics takes a lot of work and discipline. Know that you are not alone in this endeavor. We are very much interested in supporting you throughout the process of understanding and internalizing the material. Instructors, peers, and tutors are great resources to support you. Jump in with both feet, schedule your study sessions, take steps to ensure you can attend to the material sufficiently, and enjoy the journey. Good luck!

CHAPTER OVERVIEW

1: Introduction to Statistics

1.1: What is Statistics?

Learning Objectives

- Identify situations in which statistics can be misleading
- Define statistics

Introduction to Statistics

Statistics includes numerical facts and figures. For instance:

- The largest earthquake measured 9.5 on the Richter scale.
- In 2023, 83% of all adult homicide victims were male.
- in 2023, about a quarter of women of reproductive age in South Africa were HIV positive.
- By the year 2050, there will be 82 million people aged 65 and over in the United States, a 47% increase since 2022.

The study of statistics involves math and various calculations, but as a body of knowledge, statistics is built upon much more. Statistics includes theoretical frameworks that guide the formulation of questions and the data collection, analysis, and interpretation needed to answer those questions. Consider the following three scenarios where interpretations are given based on presented statistical measures. You will find that the numbers may be correct, but the interpretation may be wrong. Try to identify a major flaw with each interpretation on your own, and then check your response.

Text Exercise 1.1.1

A new advertisement for Ben and Jerry's ice cream was introduced in late May of last year. The following three months saw a 30% increase in ice cream sales; thus, the advertisement was effective.

Answer

A major flaw is that ice cream sales generally increase in the months of June, July, and August, regardless of advertisements. This effect is called a history effect and leads people to interpret outcomes as the result of one variable when another variable (in this case, the time of the year) is actually responsible.

Text Exercise 1.1.2

The more churches there are in a city, the more crime there is. Thus, churches lead to crime.

Answer

A major flaw is that both increased churches and increased crime rates can be explained by larger populations. In bigger cities, there are both more churches and more crime. This problem refers to the third-variable problem; namely, people erroneously believe that there is a causal relationship between the two primary variables rather than recognize that a third variable can cause both.

Text Exercise 1.1.3

At Harvard, the percentage of seniors that graduated with a GPA of 4.0 increased nearly 78% from 2020 to 2023. Thus, grade inflation is a real epidemic.

Answer

A major flaw is that we don't have the information that we need. What are the actual rates of occurrence? Suppose only 1% of students earned a 4.0 GPA in $2020$and $1.78\%$of graduating seniors earned such a GPA in 2023 $1.78$is 78% higher than 1. But this latter number is hardly evidence suggesting an epidemic in grade inflation. In addition, the statistic provided does not rule out the possibility that the number of 4.0 GPAs had seen dramatic fluctuations in those years due to a variety of different causes. Again, there is simply not enough information to fully understand the impact of the statistic. If you're interested, read more about the statistics of graduating seniors here.

As a whole, these examples show that "statistics" are not only facts and figures. In the broadest sense, **statistics** refers to a large range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.

1.2: Importance of Statistics

Learning Objectives

- Motivate the importance of statistical literacy in our daily lives
- Review the scientific method
- Define variable
- Define independent variable
- Define dependent variable
- Outline the basic process of statistics-based research

General Overview of Why/How We Study Statistics

Most of us are probably aware of the human desire to live a good life. In trying to achieve this goal, we must make choices based upon the world around us. To help us make these decisions, we can examine things that are consistent and predictable and use them to anticipate future events. This, paired with the knowledge that not everything in life can be classified as consistent or predictable, can help us develop tools that facilitate our growth.

What are some examples of the challenges we find in life that draw us to develop some understanding of these tools and how the world works? Perhaps a candidate for mayor claims that the rising crime rate in the city is a direct consequence of the policies his political opponent supports; how can we determine if this is correct? How do we decide who we vote for in the upcoming election? Perhaps a salesperson is trying to sell us blackout curtains, claiming that it will lower the heating bill of the house. Will we actually save money by purchasing the curtains? Will a new diet fix our health issues? Can we improve the fuel economy of our vehicles by using a different kind of fuel? Is the probability of severe weather high enough to justify canceling some event? There are many situations that require us to observe and analyze in order to make the best choice.

Having experienced the unpredictability and inconsistency found in both the natural world and our human societies, a legitimate question arises: how do we know that our understanding of the world actually represents reality? The world is complicated. Numerous factors, both seen and unseen, play into every event. People may make claims out of ignorance, or they might have ulterior, even malicious, intentions. Knowing this, how should we act? We could very easily become overwhelmed with doubts, but luckily, throughout most of our lives, we have been developing ways to address them.

The answer is quite simple and familiar. We observe people and events repeatedly throughout our lives, noting different circumstances and outcomes. We analyze our observations to form an initial conclusion. From our initial perceptions, we refine and build trust in our models by repeatedly testing and continually updating them. Eventually, our perceptions of certain individuals and events garner enough trust and consistency that we naturally rely on them. However, we are always open to additional information that may cast new light upon our previous perceptions. In most of us, this process happens naturally. Hopefully, we recognize this process as the foundation of the scientific method.

We can understand the scientific method as the result of recognizing and honing the natural process of inquiry outlined above. As we grow in our ability to analyze the world, common errors and methodological inefficiencies are identified and expunged. The scientific method begins with a set of observations that elicit some interest which then fosters the development of a research question on a particular topic. At this stage, an initial generalization or hypothesis is constructed to provide insight into an answer to the research question; the hypothesis speaks to the relationship between specific aspects of the field of interest. These specific aspects are **variables** (properties or characteristics of some event, object, or person that can take on different values or amounts). The hypothesis must be falsifiable; that is, it could be shown to be incorrect.

Once the hypothesis has been constructed, the hypothesis is tested through experimentation. When we merely observe, we cannot account for the individual influence of each variable at play. The experimental design process is one of the most important steps in the scientific method. Here, the researcher identifies all of the other variables, called **confounding variables**, that may affect the hypothesized relationship. The researcher may devise a plan to negate or control the influence of those confounding variables while systematically changing some of the variables of interest. The variables that are changed systematically by a researcher during an experiment are called **independent variables**. The **dependent variables** are the variables measured as the independent variables are manipulated. The experimental design becomes significantly more complicated as the number of independent and dependent variables increases since knowledge of how each independent variable interacts with each dependent variable would need to be

examined. For this reason, it is preferable to keep the number of independent and dependent variables to a minimum within a particular experiment. More variables can be examined in subsequent experiments.

? Text Exercise 1.2.1

High schools often prepare students for graduation by exploring career options and the various paths into those fields. A common component of such presentations is job satisfaction. In which fields are people most satisfied with their jobs? In which fields are people happiest? A young student might look at a list of careers with high satisfaction, which includes clergy, chiropractors, firefighters, nurses, and dentists (to name a few), and think that picking a career from such a list will result in living a good life. We can understand the collection of job satisfaction data as a form of experimentation. Identify independent, dependent, and confounding variable(s) and assess the connection between profession and happiness.

Answer

In job satisfaction studies, the primary variables of interest are profession and satisfaction. The researchers study particular careers, which are the "values" of the variable (profession), and then measure the dependent variable (satisfaction) as the particular careers change. This makes profession the independent variable and satisfaction the dependent variable. Many factors, such as personal values, interests, and strengths, play a major role in job satisfaction. The degrees of repetition and mindlessness in a job also play a role. These variables, and many other variables left unstated, are confounding variables. Satisfaction in career and life corresponds most directly with a person's individual values, interests, and strengths. We are unique and our vocation, our call in life, will match who we are.

The experimental design process also focuses on data collection and analysis. The researcher must determine how the independent variables will be altered consistently, how the dependent variables will be measured reliably, what analyses are appropriate for the collected data, and how these analyses test the hypothesis. Recall the definition of statistics provided earlier. Fluency with statistics facilitates this process and helps ensure that our conclusions will be meaningful, not necessarily desired, but meaningful.

Once the experimental design is finished, the experiment will be conducted, the collected data will be summarized and analyzed, and a conclusion will be made regarding the hypothesis. If the initial hypothesis was found to be false, a new hypothesis could be formed incorporating the newest findings. Alternatively, the data could align with the hypothesis, which only increases our confidence in its veracity. The scientific method encourages the sharing of experimental design and conclusions. Significant and immediate confidence can be attained in rejecting hypotheses, while confidence in the truth of hypotheses comes from repeatedly conducting experiments that support the hypotheses.

Whether we explicitly engage in the scientific method or just try to make good decisions, we routinely engage in the process of observation, generalization, testing, and updating. This process requires the collection and analysis of data with the goal of drawing further conclusions and is statistical in nature. Therefore, it behooves us to take our study of statistics seriously and to utilize it in our daily lives. It will benefit us twofold: in developing our own understanding of the world and in intelligently considering the many claims of others.

Consequently, we should understand the basic process involved in statistic-based research, whether we involve ourselves informally (in a basic inquiry of our day-to-day lives) or formally (in some important inquiry that may have an impact well beyond ourselves). The process of statistic-based research can be broadly discussed in four steps:

1. Establish a research question that is to be explored.
2. Determine appropriate subjects and needed variables to guide in producing data to address the research question. Collect such data using appropriate methods.
3. Summarize the collected data using appropriate statistical processes. Use inferential methods when needed.
4. Carefully use the summarized data to make sound, reasoned conclusion(s). The conclusion(s) should be further analyzed for practical significance, not merely statistical significance.

Although we will focus mainly on steps 3 and 4 of this process in this course, the first two steps are equally important. Establishing a quality research question, as well as determining what data is needed and how to collect that data, can be more challenging to do than the last two steps. As both producers and consumers of statistic-based research, we must be familiar with this process to understand its power and the limitations of such research.

Statistics are often presented in an effort to add credibility to an argument or advice, as can be seen in the numerous advertisements viewed daily. Many of the numbers thrown around do not represent careful statistical analysis. They can be misleading and push us into decisions that we might regret. To be an intelligent consumer of statistics, our first reflex must be to question the statistics that we encounter. We must think about the claims, the numbers, their sources, and most importantly, the procedures used to generate them.

1.3: Two Realms of Statistics- Descriptive and Inferential

Learning Objectives

- Define data
- Define descriptive statistics
- Distinguish between a sample and a population
- Define biased sample
- Introduce sample statistics and population parameters
- Define inferential statistics

Descriptive Statistics and Inferential Statistics

As we have discussed, our lives and the scientific method involve significant amounts of observation and experimentation. During both of these processes, we are gathering information. **Data** refers to information that has been collected from observation, experimentation, surveying, historical records, etc. We study the collected data to understand what is happening around us. Looking for patterns in raw data can be difficult, especially if we have many observations and measurements. In this course, we will review and develop various ways to summarize and visualize raw data. Numbers that are used to summarize and describe data are called **descriptive statistics**. In order to understand a data set sufficiently, we must use several descriptive statistics.This need will be explored in depth in Chapter 2.

? Text Exercise 1.3.1

Data and descriptive statistics are closely related to each other and are sometimes confused. For each of the following claims, identify the data and any descriptive statistics. Note that sometimes the data is implied as opposed to directly given.

1. The average score on Exam 3 was 76% for this statistics course last semester.

Answer

The data is only referenced implicitly. Since we are looking at the average score on Exam 3 from last semester, the data would consist of all scores on Exam 3 from this course last semester. 76% provides a summary of the data and is a descriptive statistic.

2. We spent $\$4160$ on groceries and household goods last year.

Answer

Again the data is only referenced implicitly. Our data consists of the expenditures from the previous year related to groceries and household goods. These costs could be found on receipts, bank records, credit card statements, or some combination. The value $\$4,160$summarizes the data by summing all the values together and is thus a descriptive statistic.

3. We have four children, aged 2, 4, 6, and 9. The oldest is 9 years old, and the average age is 5.25 years.

Answer

The data explicitly consists of the ages of the four children: $\{2,4,6,9\}$.The oldest being 9 summarizes the data by providing us with the maximum value. Both the maximum and average values are descriptive statistics. Notice that descriptive statistics can be values in the original data but do not necessarily have to be.

We collect data each and every day to help us understand the world and act accordingly, but most of the time, our interest lies beyond understanding just the collected data. We hope to generalize, to use descriptive statistics from our collected data to make a claim on a larger scale. This generalization process of extending claims to larger audiences is called **inferential statistics.** We are inferring statistics describing a data set we do not have based on a smaller data set that we do have. This process is necessary as it is impossible to collect exhaustive data for most situations and research questions. Even when collecting exhaustive data is possible, we still utilize inferential statistics to help mitigate costs while balancing accuracy.

? Text Exercise 1.3.2

Consider the difficulty of collecting all the necessary data in the following situation and the need to generalize from the possible data.

The National Election Commission has hired us to examine how U.S. citizens feel about the fairness of the voting procedures in the U.S.

Answer

To ask every single U.S. citizen how he or she feels about the fairness of the voting procedures is practically impossible. U.S. citizens live throughout the entire world. Even if all the contact information could be gathered, we could not guarantee a response whether we visit, email, or call. The time and financial costs associated would be prohibitive. Perspectives could change by the time the data collection is completed. Inferential statistics will be necessary. We will need to determine which U.S. citizens to collect data from and use that data to estimate the views of the entire country.

Populations and Samples

In inferential statistics, we draw inferences (conclusions) about large sets of data using data from a small subset of those same subjects or events. The entire data set from all subjects/events of interest is the **population**. Any smaller subset of the population data set is the **sample**. Samples are used to gain insight into the population from which it originated.

In the previous example, the population we are interested in consists of hundreds of millions of U.S. citizens. Those who we actually interviewed would constitute our sample. We would probably sample a few thousand U.S. citizens drawn from the hundreds of millions that make up the population. When choosing a sample, ensuring that one type of citizen does not have more representation than another is crucial. For example, something would be wrong with our sample if the sample happened to be made up entirely of Florida residents. A sample exclusively composed of Floridians should not be used to infer the attitudes of other U.S. citizens. The same problem would arise if the sample were comprised only of Republicans. When these types of situations occur, we say that our sample is **biased**; it over-represents or under-represents a relevant segment of the population of interest.

Inferential statistics consists of mathematical frameworks that convert information about a sample into intelligent estimations about the population from which the sample was drawn. Our estimations depend on how representative our sample is of the population. How can we ensure that our sample is a good, unbiased representation? While the task is impossible without perfect knowledge, we can address the concern by building inferential statistics around random sampling. We trust a large enough, random sample to represent different segments of society in close to the appropriate proportions and that any bias in the sample is purely by chance.

? Text Exercise 1.3.3

Consider the difficulty of collecting all the necessary data in the following situation and the need to generalize from a sample. How could we construct a random sample and estimate the value of interest?

We are interested in examining the average number of math classes taken by current graduating seniors at U.S. colleges and universities during their four years in college.

Answer

Our population consists of just the graduating seniors throughout the country. This is still a large set since there are thousands of colleges and universities, each enrolling many students. In 2022, over 2 million bachelor's degrees were granted in the United States. The cost to examine the transcript of every college senior would be prohibitive. We must construct a sample of college seniors and then make inferences to the entire population based on what we find.

To make a sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our sample was 3.2. Then we might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But, we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors or chose too many technical

institutions that have heavy math requirements. Such bad sampling would make our sample unrepresentative of the population of all seniors.

Building from this example, we mentioned that over 2 million bachelor's degrees were awarded in the United States in 2022. Since this figure describes the population, it is what we would call a **parameter**. Furthermore, we collected a sample and calculated that the average number of math classes was 3.2 per student. This figure describes the sample, referred to as a **statistic**. To summarize, a population is described by parameters, while a sample is described by statistics.

? Text Exercise 1.3.4

Identify the population and the sample, then reflect on whether the sample will likely yield the desired information.

1. A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well.

Answer

The population consists of all students in the class. The sample comprises the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those sitting in the front row tend to be more interested in the class and perform higher on tests. The sample may perform at a higher level than the population.

2. A coach is interested in how many cartwheels the average college freshman at his university can do. Eight volunteers from the freshman class stepped forward. After observing their performance, the coach concluded that college freshmen can do an average of 16 cartwheels in a row without stopping.

Answer

The population is the class of all freshmen at the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman: people who cannot do cartwheels probably did not volunteer!

? Text Exercise 1.3.5

Determine when descriptive and inferential statistics are being utilized. Assess the quality of the inference.

A quick Google Maps search showed that there were 20 Chick-fil-A restaurants open in Kansas in May of 2024. This means there were about 20 per state for a total of 1000 Chick-fil-A restaurants in the United States.

Answer

It appears that we are interested in the total number of Chick-fil-A restaurants in the United States. To guess the number that characterizes the population (the United States), a sample (Kansas) was taken, and the number of Chick-fil-A restaurants in the sample was determined. Summarizing the sample data with the number 20 would be classified as a descriptive statistic. Estimating the total number of Chick-fil-A restaurants in the United States to be $20 \cdot 50 = 1000$ belongs to the realm of inferential statistics. Descriptive statistics merely describes data, while inferential statistics makes informed guesses about what goes beyond the collected data. The inference is dubious. Kansas is a state with a relatively small population. A better sampling option would be to randomly pick more states. Indeed, if we cared about this situation, inferential statistics would be unnecessary. The desired information is readily available through Chick-fil-A itself; there were over 3000.

1.4: Sampling Methods

Objectives

- Identify biased samples
- Distinguish between methods of sampling
- Distinguish between random sampling and random assignment

Why We Sample

Sampling plays a significant role in inferential statistics. Keeping in mind that our goal is to use data from a sample to infer about the larger population, we must ensure that our sample is representative by selecting it to be sufficiently large and without any systematic biases. There are many ways to sample; some are better than others.

Simple Random Sampling

Researchers adopt a variety of quality sampling strategies. The most straightforward is **simple random sampling**. Such sampling requires every member of the population to have an equal chance of being selected for the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, choosing one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance.

? Text Exercise 1.4.1

What is the population? What is the sample? Was the sample picked by simple random sampling? Is the sample biased?

A research scientist is interested in studying the experiences of twins raised together versus those raised apart. She obtains a list of twins from the National Twin Registry and selects two subsets of individuals for her study. First, she chooses all those in the registry whose last name begins with Z. Then she turns to all those whose last name begins with B. Because there are so many names that start with B, our researcher decided to incorporate only every other name into her sample. Finally, she mails out a survey and compares characteristics of twins raised apart versus together.

Answer

The population consists of all twins recorded in the National Twin Registry. It is important that the researcher only make statistical generalizations to the twins on this list, not to all twins in the nation or world. That is, the National Twin Registry may not be representative of all twins. Even if inferences are limited to the Registry, a number of problems affect the sampling procedure we described. For instance, choosing only twins whose last names begin with Z does not give every individual an equal chance of being selected into the sample. Moreover, such a procedure risks over-representing ethnic groups with many surnames that begin with Z. There are other reasons why choosing just the $Z's$ may bias the sample. Perhaps such people are more patient than average because they often find themselves at the end of the line! The same problem occurs with choosing twins whose last name begins with B. An additional problem for the $B's$ is that the "every-other-one" procedure (called systematic sampling) disallowed adjacent names on the B part of the list from being both selected. Just this defect alone means the sample was not formed through simple random sampling.

Sample Size Matters

Recall that the definition of a simple random sample is a sample in which every member of the population has an equal chance of being selected. The sampling procedure defines what it means for a sample to be random, not the results. Random samples, especially if the sample size is small, are not necessarily representative of the entire population. For example, if a simple random sample of subjects was taken from a large enough population with an equal number of males and females, it would be about 10 times more likely that the sample consisted of 80% women if we only sampled 10 people as opposed to 20 people (4.3945% vs 0.4621%).A sample consisting of 80% women would not be representative, although the sample would be drawn randomly. Large sample sizes make it more likely that our sample is close to representative of the population. For this reason, inferential statistics takes into account the sample size when generalizing results from samples to populations. In later chapters, we will see what kinds of mathematical techniques ensure this sensitivity to sample size.

Other Sampling Methods

Our goal in constructing a sample is to arrive at a representation that yields accurate inferences regarding the population. The simplest way to guarantee that we are not systematically biased in our sampling methodology is to use a simple random sample. At times, we are aware that our population has distinct groups that differ from each other in a significantly relevant way to the topic at hand. If this were the case, we would want to ensure that each of these distinct groups would be represented in our sample. How could we guarantee appropriate representation without systematically biasing our sample?

Stratified random sampling can be an effective sampling method to guarantee the representation of different groups in a population that has natural differences. These distinct groups, known as strata, are each randomly sampled so that their sizes in the sample are proportional to their sizes in the population.

? Text Exercise 1.4.2

Suppose we were interested in views of capital punishment at an urban university. We have the time and resources to interview 200 students. The student body is diverse with respect to age; 30% of students are older people who work during the day and enroll in night courses (average age is 39), while 70% of students are younger students who generally enroll in day classes (average age of 19). It is possible that night students have different views about capital punishment than day students. How could we use stratified sampling to get a random sample?

Answer

Since 70% of the students are day students, it makes sense to ensure that 70% of the sample consists of day students. Thus, our sample of 200 students would consist of $140\,(.7 \cdot 200 = 140)$ day students chosen at random and $60\,(.3 \cdot 200 = 60)$ night students also chosen at random. The proportion of day students in the sample and in the population (the entire university) would be the same. Inferences to the entire population of students at the university would be more secure.

Simple random sampling ensures that any bias is due to random chance, but every possible sample is equally likely to occur. This means that we could end up with a sample that is difficult to collect, which makes data collection quite costly and time-consuming. We would save a lot of time and money if we could easily get a representative, random sample. In general, this isn't possible, but if we have more information about our population, we may be able to devise a better sampling strategy than simple random sampling. What would be necessary about our population if such a sampling method were to be effective? We would need the population to be well-mixed. This means that we can divide the population into sections such that each section does not have any significant differences from other sections regarding the topic at hand.

Cluster random sampling is a method that is used when a population is divided naturally into smaller groups (clusters), and each group does not have any significant differences from the others. Once these groups are created, we randomly select a set of clusters. We differentiate two types of cluster sampling based on how the clusters are studied once randomly selected. In **single-stage cluster sampling**, every member of each of the randomly selected clusters is studied. In **double-stage cluster sampling**, a simple random sample is taken from each randomly selected cluster. Double-stage cluster sampling aids efficiency and cost management, but single-stage cluster sampling is preferred since it includes more members.

Both cluster and stratified sampling divide the population into groups and select from those groups. The difference between them is that in a stratified sample, every group is selected, whereas in a cluster sample, only some of the groups are selected.

? Text Exercise 1.4.3

Suppose we are interested in voters' views regarding a school bond for the local municipal high school. Is the use of cluster sampling to obtain a random sample appropriate? If so, how could single-stage cluster sampling be implemented?

Answer

The population would be all voters registered to vote in the city. A natural partitioning of the population would be the municipal voter precincts. Since there is just one municipal high school, there would not be competition between parents of different schools vying for more money for their particular high school. We might expect tension regarding a school bond to depend on the age and presence of children. If this highlights a major difference between precincts, to cluster using them as sections would be inappropriate. While it is true that there are different types of neighborhoods, we would expect

families at different stages of life to live in each precinct. From this, there are no major differences between precincts regarding the school bond; we can implement the method.

We could randomly select 10 precincts from the total precincts in the municipality and then survey all voters in each precinct. This would be faster and much more efficient than randomly selecting voters from every precinct (stratified sampling) and randomly selecting voters regardless of any other factors (simple random sampling).

? Text Exercise 1.4.4

We are interested in determining the average height of students at our local high school. Despite being able to measure every student, we deem conducting a census not practical. Consider both methodologies (stratified and cluster sampling) for this population. Express your thoughts.

Answer

Note answers may vary. There are two natural groupings that immediately come to mind when thinking of high school students: class rank and gender. Are there major differences in heights between men and women? Yes, men tend to be taller than women on average. This means that cluster sampling by gender would be inappropriate. Similarly, there are major differences in class rank because students generally continue to grow throughout all of high school. Thus cluster sampling by class rank would be inappropriate. As such, it seems that we have identified eight natural strata (each class split by gender), meaning stratified sampling would be appropriate.

Perhaps the high school has a "homeroom" system that groups students across ages and genders into similar groups. Such homerooms could be fine candidates for clusters, depending on how they were constructed.

While our goal is to get a representative sample, the best we can do is to guarantee that any bias in the sample is due to random chance. Inferential statistics is built upon this framework. We must be wary of sampling methods that admit systematic bias. We have already encountered several of them in the course of this book: **voluntary response** (coach with cartwheels), **convenience** (students in the front row), and **systematic** (choosing every other last name starting in B). Note that in systematic it need not be every other member of the population but every k^{th} member.

? Text Exercise 1.4.5

Construct definitions for the voluntary response and convenience sampling methods. Explain how they are related and how to distinguish between them. Discuss why the methodologies produce biased samples most of the time.

Answer

Voluntary Response: A form of sampling in which a mass request is sent out or posted asking for participation in the sample. Any member of the population who receives the request and volunteers for the study will be included in the sample.

Convenience: A form of sampling in which population members are identified and selected for the sample simply because some aspect makes collecting data easier.

Both sampling methods are based on ease of access. Convenience sampling has a much broader application than voluntary response because convenience sampling may have pretty much anything as the population. Voluntary response sampling requires that the population consists of people. However, there are convenient samples of people that fail to be voluntary response. Consider a psychology professor conducting a survey of his psychology students as a sample of the student body instead of soliciting survey participants via school-wide emails. The latter is a voluntary response, while the former is merely convenience. Also, consider studying the use of turn-signals by standing at a busy intersection and counting instances over a given period of time. It is convenient to sample drivers at a single location, and participation was not voluntary. The key to remember is that voluntary response requires self-selection on the participant's part.

Voluntary response sampling generally produces biased samples because people who feel strongly, either positively or negatively, are more prone to respond to requests.

Convenience sampling generally produces biased samples because the sample is convenient for a reason; some characteristics are common among them. Thus, the subjects that do not possess that characteristic are most likely underrepresented.

In both of these methodologies, the sample produced may be representative. We cannot confirm when this is the case, and we cannot assert that any bias is due to random chance because of the previously mentioned reasons. These methods are often used, and sometimes for good reason. Doing rigorous sampling can be costly or logistically impossible. However, skepticism is justified when assessing conclusions drawn about a population based on a convenience sample.

Random Assignment in Medical Trials

In experimental research, populations are often hypothetical. For example, in an experiment comparing the effectiveness of a new anti-depressant drug with a placebo (fake treatment), there is no actual population of individuals taking the drug. In this case, a specified population of people with some degree of depression is defined and a random sample is taken from this population. The sample is then randomly divided into two groups; one group is assigned to the treatment condition (taking the drug) and the other group is assigned to the control condition (taking the placebo). This random division of the sample into two groups is called **random assignment**. Random assignment is critical for the validity of an experiment.

For example, consider the bias that could be introduced if the first 20 subjects to show up at the experiment were assigned to the experimental group and the second 20 subjects were assigned to the control group. It is possible that subjects who show up late tend to be more depressed than those who show up early, thus making the experimental group less depressed than the control group even before the treatment was administered.

In experimental research of this kind, failing to assign subjects randomly to groups is generally more serious than having a non-random sample. Failure to randomize (the former error) invalidates the experimental findings, while a non-random sample (the latter error) simply restricts the degree to which the results are generalizable.

1.5: Variables

Learning Objectives

- Define and distinguish between qualitative and quantitative variables
- Define and distinguish between discrete and continuous variables

Introduction to Variables

Recall that **variables** are properties or characteristics of some event, object, or person that can take on different values. Just as there are different types of characteristics, there are different types of variables. Since some variable types do not admit certain computations and analyses, we must pay close attention to the variables and data that we analyze.

Qualitative and Quantitative Variables

At the most general level, we can understand variables as measuring a quality (such as hair color, eye color, religion, favorite movie, gender, and so on) or a quantity (such as height, weight, age, shoe size, temperature, and so on). As such, we have the classification of **qualitative and quantitative variables**, respectively. Intuitively, we might classify variables with numbers as quantitative and variables with characteristics as qualitative. Our intuitive grasp requires further clarification when we consider variables such as race placement $\{1^{st}, 2^{nd}, 3^{rd}, \dots\}$. Just because the values that a variable takes on can be represented or encoded with numbers does not make the variable quantitative. Doing so would render this variable classification meaningless; think about why. Having an inherent order associated with the values that a variable takes on also does not make it quantitative. A **variable is quantitative** if the arithmetic difference between any two values that it takes on is well-defined and informative. While we could subtract 1 from 3 to arrive at 2. We did not gain any additional information. The inherent order already told us that 3^{rd} is two spots lower than 1^{st}; as such, race placement is qualitative. Race finish-time, on the other hand, is a quantitative variable. The arithmetic difference gives us a meaningful measurement as to how much quicker or slower one racer was compared to another. When determining if a variable is quantitative or qualitative, consider if the gaps between possible numerical values add significant meaning.

Text Exercise 1.5.1

1. Most secondary and post-secondary schools classify their students as freshmen, sophomores, juniors, and seniors. Classify class rank as a qualitative or quantitative variable. Explain your reasoning.

Answer

While commonly labeled with $9, 10, 11,$ and 12 including the associated order in class rank, that is not sufficient to warrant a quantitative variable designation. There is no informative arithmetic difference between the different values of the variable; this makes class rank a qualitative variable.

2. Credit card companies typically assign a 16 digit number to an individual to help increase transaction security. Classify customer credit card number as a qualitative or quantitative variable. Explain your reasoning.

Answer

We must ask if the arithmetic difference between credit card numbers is meaningful and informative. Given the numerical encoding, we can compute the subtraction easily enough. Is that difference informative? The credit card number does not reflect the balance, the credit limit, the age of the account, or when the latest number was assigned. When there are two cardholders for a single account, each card holder has a different credit card number. It appears that there is no information gained in knowing the arithmetic difference. The numbers just serve as identifiers to an account; therefore, customer credit card number is qualitative.

3. We mark the passage of time with the month, day, and year. Classify the date as a qualitative or quantitative variable. Explain your reasoning.

Answer

At first glance, values like September 20, 2019 and October 25, 2021 look like they cannot be subtracted. We can count how many days lie between these two values (766 days), and hence the gap between values carries significant meaning. A well-defined arithmetic difference may be difficult write down (it can be done), but intuitively we understand that differences mean how much time has passed between the dates. This is informative. Thus, date is a quantitative variable.

Discrete and Continuous Variables

Quantitative variables can be further classified by the possible values they take on. As quantitative variables, we know that these values in their essence are numbers, but more can be told. Consider the number of people who are in attendance at a FHSU football game. The number fluctuates from game to game, but we know that the possible values are $0, 1, 2, 3, 4, 5, 6, 7, \ldots, 6362$ (Lewis Field's seating capacity). Now consider how long it takes for a football game to finish, denoted with the variable t. There must be four quarters, each with at least 15 minutes of game time and an intermission that is supposed to be at least 20 minutes. If we then consider the play clock, injuries, official reviews, timeouts, overtime, delays, and plays running past time, we would say that the time it takes for a football game to finish could be any number of minutes with 80 minutes as a minimum. For example, it could finish after 83.00000001 minutes, 92.475920 minutes, 92.4759202 minutes, or even $102.1340294478\ldots$ minutes. We could express the possible values as an interval ($\{t \geq 80\}$ or alternatively $[80, \infty)$). If we look at game attendance on the other hand, we could not express the possible values as an interval (one cannot have 7.5 people in attendance); for every pair of distinct possible values, there is a number between them that is not a possible value. Herein lies the distinction between discrete and continuous variables.

A quantitative variable is **continuous** if it can take on any numerical value in some interval of real numbers. A quantitative variable is **discrete** if it is not continuous. A common way to describe a discrete variable is to say that there are gaps between all the possible values that the variable takes on.

? Text Exercise 1.5.2

1. Consider the amount of U.S. currency stored in bank accounts. Classify this variable as a qualitative or quantitative variable. If quantitative, further classify the variable as discrete or continuous. Explain your reasoning.

Answer

The amount of money stored in bank accounts is quantitative because the differences between account balances indicate how much more or less one account has compared to another. The smallest unit of U.S. currency is the penny ($\$0.01$). This implies that there must be a gap between all possible values (one cannot have $\$5.7365$) and that the set of possible values does not contain an interval. Thus U.S. currency in a bank account is a discrete quantitative variable.

2. Consider the floors on an eighteen-story apartment building labeled in order $\{1^{st}, 2^{nd}, 3^{rd}, \ldots, 18^{th}\}$. Classify this variable as a qualitative or quantitative variable. If quantitative, further classify the variable as discrete or continuous. Explain your reasoning.

Answer

The arithmetic difference between two floors returns the same information as the inherent ordering. This variable is thus qualitative. As such, the variable does not get classified as discrete or continuous.

3. Consider the number of floors in apartment buildings. Classify this variable as a qualitative or quantitative variable. If quantitative, further classify the variable as discrete or continuous. Explain your reasoning.

Answer

The values that this variable can take on are $\{1, 2, 3, 4, \ldots\}$. The arithmetic difference between the number of floors indicates the difference in the number of floors between the different apartment buildings. It is therefore, a quantitative variable. The number of floors are counted using whole numbers; therefore, there are gaps between the possible values that the variable takes on. This makes the variable discrete.

4. Consider the heights of adult females. Classify this variable as a qualitative or quantitative variable. If quantitative, further classify the variable as discrete or continuous. Explain your reasoning.

Answer

The arithmetic difference between the heights of two adult females reveals the height disparity between them. Height of adult females is a quantitative variable. The heights could take on any value in an interval (given any two distinct height values there are other possible height values between them) and is therefore continuous.

The classification of discrete and continuous variables focuses on the possible values that a variable takes on. As you may have considered in the previous exercise, there can be a disparity between the values a variable may take on and our ability to measure them. If we based our classification on our ability to measure, every variable would be discrete. If we measured heights accurately enough, any number in some interval could potentially be observed. In principal, someone's height could be 72.65787652998736 inches, despite the fact that we would rarely take the effort to measure that accurately. Contrast this with measuring the number of people in a population; no matter how accurately we measure, we know that it is not possible to obtain a value of 800.78 or 110.2. Thus, the distinction between discrete and continuous variables lies in the possible values that could theoretically be, not the values that may be measured in practice. We are finite, limited beings, and that is good to keep at the forefront of our minds. Remember that we are seeking to understand the reality of the world around us and act accordingly.

1.6: Levels of Measurement

Learning Objectives

- Define and distinguish among nominal, ordinal, interval, and ratio levels of measurement
- Give examples of errors that can be made by failing to understand the proper use of measurement levels

Types of Measurement

When we have a particular topic of interest and want to further our understanding, we need to collect appropriate data through some sort of measurement or observation. Exactly how the measurement is carried out depends on the variable of interest. To measure the time taken to respond to a stimulus, we might use a stop watch. Stop watches are of no use, of course, when it comes to measuring someone's attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like "very favorable," "somewhat favorable," etc.). For a variable such as "favorite color," we can simply note the color-word (like "red") that the subject offers. Although measurements can differ in many ways, they can be classified using a few fundamental categories. These categories are called levels of measurement because each one is contained in the previous one. Just like we have categories of animal, mammal, canine, and dog, each referring to a differing level of specificity, so too do we have categories of nominal, ordinal, interval, and ratio. We'll start with the broadest category, nominal, and work our way down to the most specific category, ratio.

Nominal Scale

When we simply name or categorize responses, we are measuring on a nominal scale. Gender, handedness, favorite color, and religion are examples of variables measured on a nominal scale. The essential point about nominal scales is that they categorize without implying any ordering among the responses. For example, when classifying people according to their favorite color, there is no measurement sense in which green is placed "ahead of" blue. Responses are merely categorized. Nominally scaled data embody the lowest level of measurement.

Ordinal Scale

A researcher wishing to measure consumers' satisfaction with their microwave ovens might ask them to specify their feelings as either "very dissatisfied," "somewhat dissatisfied," "somewhat satisfied," or "very satisfied." The items in this scale are ordered, ranging from least to most satisfied. This is what distinguishes ordinal from nominal scales. Unlike nominal scales, ordinal scales allow comparisons because there is a meaningful order to the measurement values. For example, our satisfaction ordering makes it meaningful to assert that one person is more satisfied than another with their microwave ovens.

On the other hand, ordinal scales fail to capture important information that will be present in the subsequent levels we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels. In our satisfaction scale, for example, the difference between the responses "very dissatisfied" and "somewhat dissatisfied" is probably not equivalent to the difference between "somewhat dissatisfied" and "somewhat satisfied." Nothing in our measurement procedure allows us to determine whether the two differences reflect the same difference in psychological satisfaction. Indeed, even if we had two pairs of observations each with "very dissatisfied" and "somewhat dissatisfied" ratings, we could not determine whether the differences in these ratings are truly the same.

What if the researcher had measured satisfaction by asking consumers to indicate their level of satisfaction by choosing a number from one to four? Would the difference between the responses of one and two necessarily reflect the same difference in satisfaction as the difference between the responses two and three? The answer is No. Changing the response format to numbers does not change the meaning of the scale. We still are in no position to assert that the qualitative difference between 1 and 2 (for example) is the same as 3 and 4.

Text Exercise 1.6.1

Classify the following variables based on their nominal or ordinal level of measurement. Explain.

1. Eye color

Answer

Since there is no inherent, meaningful order to eye color (as normally denoted green, blue, hazel, and brown), eye color is nominal.

2. BMI weight type (underweight, healthy, overweight, obese, severely obese)

Answer

There is an inherent, meaningful order to BMI weight types, each subsequent value (as listed) indicates an increasing BMI. Hence BMI weight type is ordinal.

3. Shirt size (S, M, L, XL)

Answer

There is an inherent, meaningful order to shirt sizes, each subsequent value indicates a larger or smaller shirt (depending on the ordering). Hence shirt size is ordinal.

4. Phone number

Answer

Phone numbers are granted based on availability and hence have no meaningful order. Thus, phone number is nominal.

Interval Scale

Notice that qualitative data is nominal or ordinal. If we have quantitative data; that is, data with meaningful differences in values, then we have some important distinctions to make. Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).

To differentiate interval scale from our next level of measurement, we note that on interval scales a measurement of 0 does not represent the absence of some quantity. On interval data, 0 is artificially or arbitrarily defined and is not intrinsically meaningful to the quantity being measured. The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label 0 is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. This is important because if 0 doesn't mean "nothing," then it is not meaningful to divide nor multiply. For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees Fahrenheit; no interesting physical property is preserved across the two ratios. For this reason, it does not make sense to say that 80 degrees Fahrenheit is "twice as hot" as 40 degrees Fahrenheit.

One way that we mark the passage of time is by using the designation of year. Citizens of the United States remember the year 1776 as the year the Declaration of Independence was signed. Jews remember the year 70 as the year the temple was destroyed. The first recorded Olympic games occurred in 776BC. The differences between these years are meaningful; the Declaration of Independence occurred 1706 years after the Romans razed the temple and 2552 years after the first recorded Olympic games. So the year designation of time, must be at least of interval level. While there is a year 0, it does not represent the absence of passage of time; therefore, the year is measured on the interval level.

Ratio Scale

As the highest level of measurement, ratio scales allow the most varied statistical analyses. The ratio scale of measurement is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured. Like a nominal scale, the ratio scale provides a name or category for each object (the numbers serve as labels). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference between values on the scale has the same meaning; however, with the ratio scale, the same ratio between values on the scale also carries the same meaning. This is the ratio scale's defining characteristic.

The Fahrenheit scale for temperature has an arbitrary zero point and is, therefore, not a ratio scale. However, 0 on the Kelvin scale is absolute zero (total absence of kinetic energy), making this a ratio scale. For example, if one temperature is twice as high as

another as measured on the Kelvin scale, then that temperature has twice the kinetic energy of the other temperature. Therefore, it does make sense to say 80 K is twice as hot as 40 K.

Another example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 55 cents, etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money. Since money has a true zero point, we can say that someone with 50 cents has twice as much money as someone with 25 cents.

Ratio scales are very common. Most quantities of scientific interest tend to be ratio: distance, speed, weight, mass, pressure, volume, area, energy, population; these are all variables measured on ratio scales.

? Text Exercise 1.6.2

Define a variable on the ratio scale that can take on both negative and positive numbers.

Answer

Note answers may vary. Since ratio scales have "zeros" that carry meaning relative to what we are actually measuring, we might think that ratio scales cannot change signs. This conclusion might be intuitive, but unfortunately the conclusion is false. Vector measurements, measurements that include a direction, can have both positive and negative values while remaining on the ratio scale. Consider displacement, velocity, and acceleration to name a few. Electrical charge is an example of a non-vector, ratio scale which can be negative or positive. Money in an account is another example, as negative values represent debt, but 0 still means no money in the account.

Consequences of Levels of Measurement

Why are we so interested in the type of scale that measures a variable? The crux of the matter is the relationship between the variable's level of measurement and the statistics that can be meaningfully computed with that variable. For example, consider a hypothetical study in which 5 children are asked to choose their favorite color from blue, red, yellow, green, and purple. The researcher codes the results as follows:

Table 1.6.1: Guide for Encoding Colors as Numbers

Color	Code
Blue	1
Red	2
Yellow	3
Green	4
Purple	5

This means that if a child said her favorite color was "Red," then the choice was coded as 2, if the child said her favorite color was "Purple," then the response was coded as 5, and so forth. Consider the following hypothetical data:

Table 1.6.2: Favorite Colors and Code from Sample of 5 Children

Subject	Color	Code
1	Blue	1
2	Blue	1
3	Green	4
4	Green	4
5	Purple	5

Each code is a number, so nothing prevents us from computing the average code assigned to the children. The average happens to be 3, but you can see that it would be senseless to conclude that the average favorite color is yellow (the color with a code of 3). Such nonsense arises because favorite color is a nominal scale, and taking the average of its numerical labels is like counting the number of letters in the name of a snake to see how long the beast is.

In a similar fashion, does it make sense to compute the average of numbers measured on an ordinal scale? Different fields of study might answer this question differently; after all, reviews online often give an average rating from 1 to 5 stars and surveys rating 0 (strongly disagree) to 10 (strongly agree) are often summarized with an average rating. Suppose an individual ran 4 races and placed 1^{st}, 2^{nd}, 3^{rd}, and 4^{th}. We could say the average placement was 2.5^{th}, but what does that mean? Suppose another individual placed 3^{rd}, 3^{rd}, 1^{st}, and 3^{rd} in the same races to also have an average placement of 2.5^{th}. Is it meaningful to say these two runners are, on average, equally fast? We cannot conclude that. Perhaps the second runner lost the first two races by a large margin, but the second two races were neck and neck. The problem is that we only know who was faster, we don't know by how much. We, therefore, recommend avoiding such a practice and considering results based on averaged ordinal data with a careful eye.

To conclude such a discussion, once we attain the levels of interval and ratio, computing the average of our data is well-defined and meaningful. As we move forward, always be aware of what descriptive statistics are possible given the consequences of the level of measurement at hand.

CHAPTER OVERVIEW

2: Descriptive Statistics

2.1: Descriptive Statistics and Distributions

Learning Objectives

- Define and distinguish between statistics and parameters
- Define and calculate proportions
- Introduce and distinguish between frequency and relative frequency distributions
- Introduce and use summation notation
- Introduce grouped distributions
- Introduce bar charts and histograms
- Identify the skew of a distribution

Section 2.1 Excel File (contains all of the data sets for this section)

Population Parameters and Sample Statistics

We gain a better understanding of the world around us by collecting and analyzing data. Recall that, most of the time, it is not possible or practical to collect all the data around a certain topic. For this reason, we often rely on inferential statistics to make informed guesses about the population using data from a random sample. It is important for our analyses to differentiate facts about the population from facts about a sample. Facts about sample data are called **statistics** while facts about populations are called **parameters**. It is common, but not universal, to use Greek letters (such as μ, σ) when referring to population parameters and Latin letters (such as $\bar{x}, s$) when referring to sample statistics. Our first example highlights one of the exceptions to this practice.

Both sample statistics and population parameters fall under the umbrella of descriptive statistics; they are numbers that are used to summarize and describe data. A commonly used descriptive statistic is the proportion. A **proportion** is the percentage of observations that have a certain characteristic. Many important issues rely on estimating proportions. What proportion of customers are satisfied with our services? What proportion of people who take some medicine experience negative side effects? What proportion of voters support this political candidate? The symbol for population proportion is p, and the symbol for sample proportion is $\hat{p}$ (read: p-hat). For example, suppose a survey was given out a month before a local election. Out of the 100 people surveyed, 54 supported a particular candidate. However, on the actual election day, that candidate only got 48% of the votes. We would say $\hat{p} = \frac{54}{100} = 0.54$ and $p = 0.48$.

Text Exercise 2.1.1

A husband (Adam) and wife (Betsy) have three children (Cathy, Damon, and Erin). Adam, Betsy, and Cathy wear glasses. We are interested in studying this particular family.

1. Compute the population proportion p of family members that wear glasses.

Answer

To find the percentage of family members that wear glasses, we need to know the total number of family members (the population size N) and the number of family members that wear glasses (the number of observations with the characteristic x). $N = 5$ and $x = 3$. Thus

$$p = \frac{x}{N} = \frac{3}{5} = 0.6$$

Thus 60% of the family wears glasses.

2. Construct the different samples and show that there is no sample such that $p = \hat{p}$.

Answer

To show that there is no sample such that $\hat{p} = p$, we must consider all possible samples from the population. The sample size n could be any number $\{1, 2, 3, 4\}$. If $n = 1$, then we could have someone with glasses or someone without glasses. Thus

$$\hat{p} = \frac{x}{n} = \begin{cases} \frac{0}{1} = 0 \\ \frac{1}{1} = 1 \end{cases}$$

If you are unfamiliar with this notation, we are saying that $\hat{p}$ could be 0 or 1. If $n = 2$, then we could have $0, 1$, or 2 people with glasses. Thus

$$\hat{p} = \begin{cases} \frac{0}{2} = 0 \\ \frac{1}{2} = 0.5 \\ \frac{2}{2} = 1 \end{cases}$$

If $n = 3$, then we could have $1, 2$, or 3 people with glasses (since there are only two people without glasses). Thus

$$\hat{p} = \begin{cases} \frac{1}{3} = 0.\overline{3} \\ \frac{2}{3} = 0.\overline{6} \\ \frac{3}{3} = 1 \end{cases}$$

If $n = 4$, then we could still only have 2, or 3 people with glasses. Thus

$$\hat{p} = \begin{cases} \frac{2}{4} = 0.5 \\ \frac{3}{4} = 0.75 \end{cases}$$

We notice that none of the possible $\hat{p}$ values match the calculated p value.

3. Suppose they had another child, Frank, show that it is now possible to have a sample such that $p = \hat{p}$

Answer

With the addition of Frank, $N = 6$ and $x = 3$ or 4 since we do not know whether Frank wears glasses or not. Thus

$$p = \begin{cases} \frac{3}{6} = 0.5 \\ \frac{4}{6} = 0.\overline{6} \end{cases}$$

Since we are showing that it is possible, finding particular samples will be sufficient. If $n = 2$, we could have someone with glasses and someone without glasses, and $\hat{p} = \frac{1}{2}$. If $n = 3$, we could have 2 people with glasses and 1 person without glasses, and $\hat{p} = \frac{2}{3}$. Thus, in this situation it is possible to have a sample such that $p = \hat{p}$.

Notice all of the proportions calculated throughout the example fell between 0 and 1. Proportions are the percentage of observations that have a certain characteristic; it is impossible to have negative numbers of observations just as it is impossible to have more observations with a certain characteristic than the total. Knowing what values are possible helps us identify when we make mistakes. We must always ask if our results are reasonable.

Distributions

Getting a firm grasp on a set of data generally requires several descriptive statistics and a method of visualization. Two very different data sets may have the same values for certain descriptive statistics while differing for others. A good place to start is to see how the data is distributed. We will build our understanding through examples.

A recently purchased bag of Plain M&M's contained six different colors of candy. A quick count showed that there were 55 M&M's: 17 brown, 18 red, 7 yellow, 7 green, 2 blue, and 4 orange. Consider Table 2.1.1 below.

Table 2.1.1: Frequencies and Relative Frequencies of Sampled M&M's

Color	Frequency	Relative Frequency
Brown	17	$\frac{17}{55} \approx 0.309$
Red	18	$\frac{18}{55} \approx 0.327$
Yellow	7	$\frac{7}{55} \approx 0.127$
Green	7	$\frac{7}{55} \approx 0.127$
Blue	2	$\frac{2}{55} \approx 0.036$
Orange	4	$\frac{4}{55} \approx 0.073$

This table describes both the **frequency distribution** and the **relative frequency distribution** of M&M's by color. The colors form what we call **classes**. Since, every M&M must belong to a class, we say the classes are **exhaustive**. Since any particular M&M cannot be classified in more than one class, the classes are **mutually exclusive**. These two properties are important to guarantee that we count each observation only once. Distributions are often shown graphically with **bar graphs** as in Figure 2.1.1.

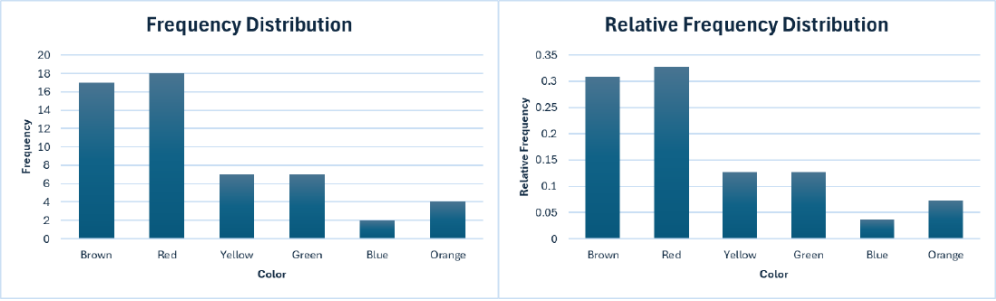


Figure 2.1.1: Frequency and Relative Frequency Distributions of 55 M&M's.

Notice how the two distributions show essentially the same information about where the data falls. Most of the M&M's were either brown or red. Yellow and green appeared equally often. Blue occurred the fewest number of times. What is the difference between the two distributions? If we looked closely at the table, we saw how the relative frequency column was computed; the frequency of the particular color was divided by the total number of M&M's. Hopefully, this reminds us of the computation for proportions. The **relative frequency** is simply the percentage of observations that have the characteristic defining the class.

Text Exercise 2.1.2

Using Table 2.1.1, determine the sum of all the relative frequencies in the relative frequency distribution of M&M's. Explain why your result must be true for all relative frequency distributions.

Answer

$\frac{17}{55}+\frac{18}{55}+\frac{7}{55}+\frac{7}{55}+\frac{2}{55}+\frac{4}{55}=\frac{55}{55}=1$. This is true for all relative frequency distributions. Since classes must be exhaustive and mutually exclusive, each observation must be in one and only one class. This means that the sum of all frequencies must add up to the total number of observations. Now relative frequencies are just frequencies divided by the total number of observations. In adding up all the relative frequencies, we could factor out the total number of observations in the denominator $\frac{1}{55}(17+18+7+7+2+4)=\frac{55}{55}=1$ and arrive at the sum of all frequencies divided by the total number of observations which is 1.

Note: Summation Notation

The explanation in the previous exercise is a bit tedious: writing down all 6 numbers repeatedly is inconvenient, and imagine repeating this exercise if there were 100 colors! Mathematicians have developed a notation to help express such arguments and computations easily. We call it summation notation. The capital Greek letter sigma $\sum$ is what we use to denote a summation. We then name all the terms that we are adding together. In our M&M's example, there were six classes and we were interested in the frequency of each class. We might refer to our frequencies by row (from the top) as f_i for $i \in \{1,2,3,4,5,6\}$. The following expression is the sum of all the relative frequencies of M&M's:

$$\sum_{i=1}^{6} \frac{f_i}{n} = \frac{f_1}{n}+\frac{f_2}{n}+\frac{f_3}{n}+\frac{f_4}{n}+\frac{f_5}{n}+\frac{f_6}{n}=1$$

For instance, $f_1 = 17$ and $f_2 = 18$. The following expression is the sum of all the relative frequencies of M&M's:

$$\sum_{i=1}^{6} \frac{f_i}{n} = \frac{17}{55}+\frac{18}{55}+\frac{7}{55}+\frac{7}{55}+\frac{2}{55}+\frac{4}{55}=\frac{55}{55}=1$$

Note that the $i = 1$ at the bottom tells us where to begin (the first class in this case) and the 6 up top tells us to add each subsequent term through that index value (the sixth class in this case).

With this notation, we can clean up our argument from the previous exercise. Suppose that we have a sample of size n with k classes. If we let f_i be the frequency in the i^{th} class, then $\sum_{i=1}^{k} f_i = n$. The sum of all of the relative frequencies would be

$$\sum_{i=1}^{k} \frac{f_i}{n} = \frac{f_1}{n}+\frac{f_2}{n}+\frac{f_3}{n}+\ldots+\frac{f_{k-1}}{n}+\frac{f_k}{n}=\frac{1}{n}\sum_{i=1}^{k} f_i=\frac{n}{n}=1$$

We encourage the reader to be familiar with summation notation, as it will be used throughout the text. If not understood, many equations given later may be confusing. It would be good to practice using the notation now while the examples are relatively simple.

Text Exercise 2.1.3

Using the frequencies from our bag of M&M's, compute the following summations:

1. $\sum_{i=3}^{5} f_i$

Answer

$\sum_{i=3}^{5} f_i = f_3+f_4+f_5 = 7+7+2 = 16$

2. $\sum_{i=2}^{4} 2f_i$

Answer

$$\sum_{i=2}^{4} 2f_i = 2f_2 + 2f_3 + 2f_4 = 2(f_2 + f_3 + f_4) = 2(18 + 7 + 7) = 2(32) = 64$$

3. $\sum_{i=1}^{3} f_i^2$

Answer

$$\sum_{i=1}^{3} f_i^2 = f_1^2 + f_2^2 + f_3^2 = 17^2 + 18^2 + 7^2 = 289 + 324 + 49 = 662$$

4. $\left(\sum_{i=1}^{3} f_i\right)^2$

Answer

$$\left(\sum_{i=1}^{3} f_i\right)^2 = (f_1 + f_2 + f_3)^2 = (17 + 18 + 7)^2 = 42^2 = 1764$$

5. $\sum_{i=1}^{6} f_i f_{7-i}$

Answer

$$\sum_{i=1}^{6} f_i f_{7-i} = f_1 f_6 + f_2 f_5 + f_3 f_4 + f_4 f_3 + f_5 f_2 + f_6 f_1 = 17 \cdot 4 + 18 \cdot 2 + 7 \cdot 7 + 7 \cdot 7 + 2 \cdot 18 + 4 \cdot 17 = 68 + 36 + 49 + 49 + 36 + 68 = 306$$

Text Exercise 2.1.4

One class of 30 students had a 10 point assignment. The student scores (raw data) were tabulated in the following set. Use the set to construct a frequency distribution in a table.

$$\{3, 4, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10\}$$

Answer

Looking at the set of scores, our classes can consist of $\{3, 4, 5, 6, 7, 8, 9, 10\}$All that is left is to count the number of observations in each class and put them in a table.

Table 2.1.2: Grouped Frequency Distribution

Student Score	Frequency
3	1
4	1
5	3
6	5
7	5
8	7
9	5
10	3

The distribution shown in Figure 2.1.1 concerns just the one bag of M&M's. We might expand our study to the distribution of colors for all regular Plain M&M's. Only the manufacturer of M&M's could provide this sort of information, but they do not tell us exactly how many M&M's of each color were ever produced. Instead, they only report the relative frequencies. See Figure 2.1.2.

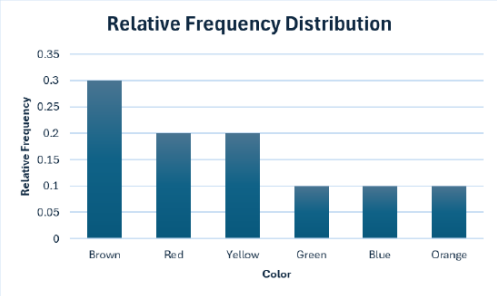


Figure 2.1.2: Distribution of all M&M's.

Notice that the relative frequency distributions in Figures 2.1.1 and 2.1.2 are not identical. Figure 2.1.1 portrays the distribution in a sample of 55 M&M's. Figure 2.1.2 shows the distribution of all M&M's. Chance factors involving the machines used by the manufacturer introduce random variation into the different bags produced. Some bags will have a distribution of colors that is close to Figure 2.1.2; others will be much different. This reinforces an important concept; sample data most often do not produce the exact same distribution/measures as what is happening in population data. We must remember this important concept to interpret our findings properly and to draw appropriate conclusions.

Distributions of Continuous Variables

Using the color of M&M's for classes seems natural, but that was not the only set of classes that could have been used. We could sort the colors as warm (red, yellow, and orange) or cool (brown, green, and blue). We could use any variable regarding M&Ms as a basis for our classes, such as weight. If we did not have a precise enough scale to differentiate weights between individual candies, weight might not been helpful. On the other hand, if our scale was too exact, we might not have had many measurements that were precisely the same. In either case, our frequency distribution would have been uninformative. Having precise measurements is a good thing; we do not want to "fix" this issue by settling for lower-quality data, but instead, we can address the problem by how we define classes. Rather than having a singular value determine our classes, we define them using a range of values. The classes must still be exhaustive and mutually exclusive. When we group values to build classes, we describe the frequency and relative frequency distributions as **grouped distributions** (this term applies as long as various values are grouped together to form classes regardless of whether we have discrete or continuous data).

The data shown in Table 2.1.3 are the times (in milliseconds) it took to move the mouse over a tiny target in a series of 20 trials. The times are sorted from shortest to longest. The variable "time to respond" is a continuous variable. With time measured so precisely, no two response times were the same; creating a grouped frequency distribution is in order.

Table 2.1.3: Response Times

568	645	720	824
577	657	728	825
581	673	729	865
640	696	777	875
641	703	808	1007

Table 2.1.4 shows one of many possible choices we could have made for a grouped frequency distribution of these 20 times. It is important to note that there is flexibility in the number of classes and where they start. Constructing multiple tables and graphs with various class numbers, sizes, and starting places helps us understand the data. We can select the most enlightening version.

Table 2.1.4: Grouped frequency distribution

Class	Class	Frequency
500 to 600	$(500, 600]$	3
600 to 700	$(600, 700]$	6
700 to 800	$(700, 800]$	5
800 to 900	$(800, 900]$	5
900 to 1000	$(900, 1000]$	0
1000 to 1100	$(1000, 1100]$	1

Notice that the classes cover all numbers from 500 to 1100, and each has the same length. To ensure that classes are mutually exclusive, we need to clarify where $600, 700, 800, 900$, and 1000 belong. While getting such a value is unlikely, paying attention to the details is essential. We set the lower bounds to be exclusive and the upper bounds to be inclusive. For example, an observation of precisely 900 milliseconds would not be assigned to $900 - 1000$ but rather be assigned to $800 - 900$. Nothing is objective about this choice; we could have decided to do the reverse. What's important is that it is consistent across all classes and that the classes remain mutually exclusive.

Note: Interval Notation

We use interval notation as a way to describe a continuous set of numbers and how we include (or not include) the endpoints.

The use of (a, b) implies that we are including all possible numbers between a and b, but we would not include either number a or b. We can choose numbers very close to a and b but never equal to a and b.

The use of $[a, b]$ implies that we are including all possible numbers between a and b including the endpoints a and b.

We can also use both in a single interval; $(1000, 1100]$says we are taking all of the numbers from 1000 up to 1100 including the number 1100 but not 1000.

As with our previous distributions, grouped frequency distributions can be portrayed graphically. Figure $2.1.3$ shows a graphical representation of the grouped frequency distribution in Table $2.1.3$. Notice there are no longer gaps between all of the bars. We do this to emphasize that this is a grouped distribution of a continuous quantitative variable. The graph of a frequency or relative frequency distribution of a continuous quantitative variable is called a **histogram** (note that some statisticians extend this name to the graphs of distributions of quantitative variables in general, others just to grouped distributions).

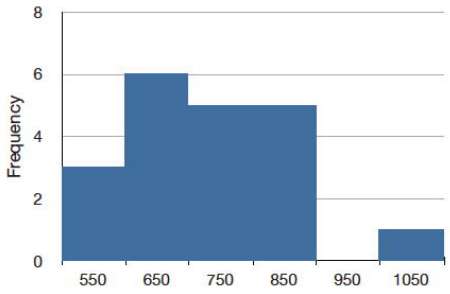


Figure $2.1.3$: A histogram of the grouped frequency distribution shown in Table $2.1.3$. The labels on the horizontal axis are the middle values of the range they represent.

Shapes of Distributions

The order that colors were presented in our frequency distribution for the M&M's did not matter, because color is a nominal variable. When we examine variables on the ordinal, interval, or ratio scales, we construct the distributions following the natural order. If we have a quantitative variable (on interval or ratio scale), the meaningful arithmetic differences in values allow us to describe the distribution by its general shape through a graph. We must be careful when describing the distribution (graph) of grouped data because the shape depends on how the classes were defined. We will develop a greater ability to describe distributions; for now, we will focus on three descriptions of bar graphs and histograms: symmetric, positively skewed, and negatively skewed.

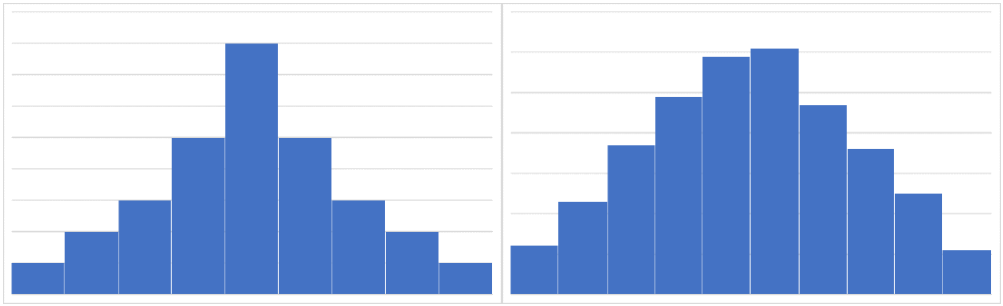


Figure $2.1.4$: Histograms of two hypothetical distributions with varying degrees of symmetry

The graphs of the distributions shown in Figure $2.1.4$, are **symmetric**; if we folded each graph in half, the two sides would match. The histogram on the left is perfectly symmetric; perfect symmetry is not likely to occur using experimental data. The histogram on the right is not perfectly symmetric (see how the four central bars would not match across the middle), but this is closer to what we would expect from observational data coming from a symmetric variable.

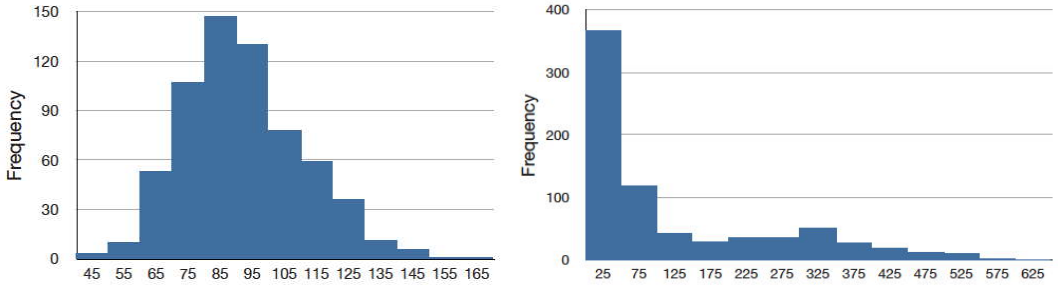


Figure 2.1.5: Two histograms (scores on a psychology test and 1974 MLB salaries (in thousands of dollars)) with varying degrees of positive skew.

Figure 2.1.5 shows two histograms that are not symmetric. Notice the ends of the graphs (called tails) in the positive direction extend further than the tails in the negative direction. A graph of a quantitative variable (bar graph or histogram) with the longer tail extending in the positive direction is said to be **positively skewed** or skewed to the right. A graph of a distribution can be **negatively skewed** or skewed to the left. These graphs have the tails in the negative direction extending further than the tails in the positive direction.

2.2: Using and Understanding Graphs

Learning Objectives

- Assess graphs for information and quality

Introduction to Graphs

Florence Nightingale $(1820 - 1910)$, although primarily known as a nurse, analyzed data in the service of her patients. She was the first woman to be a fellow of the Royal Statistical Society. She was one of the first nurses to use graphical representations to illustrate the causes of mortality. She created the graph below to share the true cause of British soldier mortality during the Crimean War clearly.

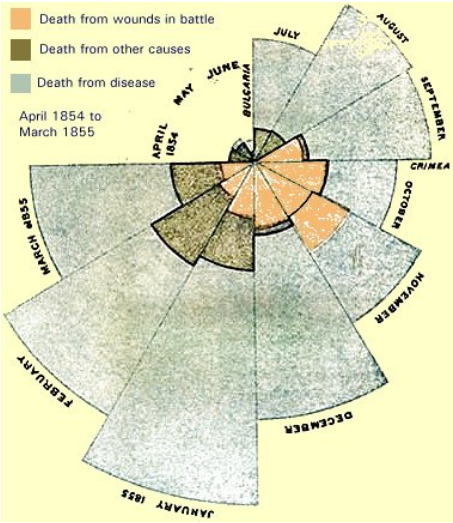


Figure 2.2.1: Polar-area diagram for British soldier mortality from April 1854 through March 1855

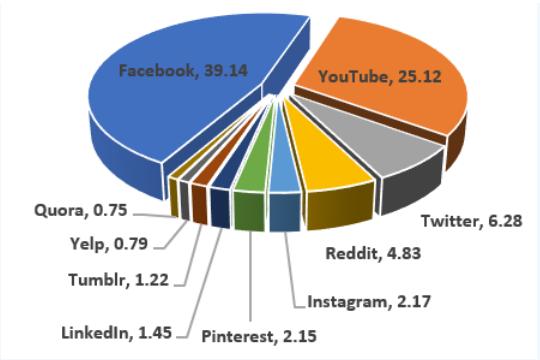
From the graph, we know that causes of mortality are broken into three classifications: wounds in battle, disease, and other causes. We can conclude relatively quickly that being wounded in battle was not the leading cause of death; disease caused most of the deaths. When Nightingale arrived, the conditions of the military hospitals were awful, and more people died from disease than their wounds. Once her sanitation practices were implemented, the mortality rate dropped from 42.7% to 2.2%. The question we must address is: how do we trust that the graph relays the truth?

We were not provided any information about the construction of the graph; please be careful of graphs without enough explanation. We guessed that the number of mortalities was related to the slice size for any given month. This graph has two components to size: the radius and the area. By which measure are we making a comparison? Most of us naturally make mental comparisons based on the area of each slice. If the graph were constructed using the radius to indicate the number of mortalities, we could still tell which months had more deaths, but we would misjudge the magnitude of these differences. Nightingale constructed the graph so that the areas represented the amount of mortalities; this information is necessary for quality interpretation and understanding.

Graphs and charts are excellent ways to share information quickly with a larger audience. In this section, we will look at several different types of graphs. However, our goal is to identify certain types of graphs to become informed consumers of information and critical thinkers actively engaged in the world around us.

Text Exercise 2.2.1

The pie graph below shows the percentage of visits to social media sites in 2017. There are various issues with this particular pie graph; identify some of the problems.



Answer

Pie graphs effectively display the relative frequencies of a small number of categories. However, pie graphs with a large number of categories are not recommended. This example uses too many categories.

Pie graphs are most commonly used to compare parts to the whole. This example has two components working against this comparison. The pie slices are not touching and the visual comparison of areas is skewed because of the three-dimensional formatting. When adding the third dimension to a graph, it must add information and not hide important information.

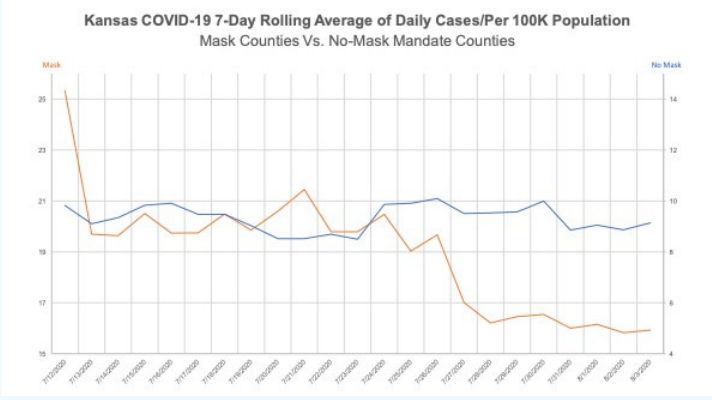
Furthermore, the numbers here do not add to 100. This means that the number given does not represent the relative size of the slice of the pie. For example, Facebook appears to take up nearly half of the pie, but the number given is 39.14. This makes it seem as if the websites listed here account for all the websites that see a relatively non-negligible amount of visits. There should be another slice, perhaps labeled "other," which takes up 16.1% of the pie.

Pie graphs are most helpful in communicating relative sizes or proportions of a few categories. They are not useful if there are numerous categories or if one wants to communicate absolute, not relative, measures. Note that relative measures are more likely to be misleading when the sample size is small.

A **time series graph** consists of the measurement of the same variable of the same subject taken over regular time intervals for a given period. Time series data frequently occurs in contexts where variables change: stock market prices, national economic figures, population tracking, etc. During the COVID-19 pandemic, time series graphs were frequently used to communicate transmission and mortality rates. Such graphs can serve as natural and valuable visual aids, but they can be misleading if not constructed appropriately.

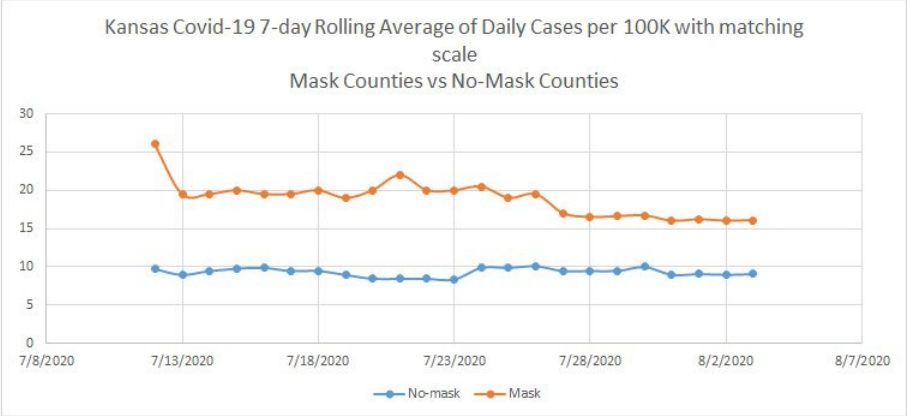
? Text Exercise 2.2.2

In August 2020, Dr. Norman, the Secretary of the Kansas Department of Health and Environment, used the following graph to show the number of COVID-19 cases in counties with mask mandates vs counties without mask mandates. Study the graph and see what conclusions can be made. Are there any issues?



Answer

First, notice that the scales on each side of the graph represent the same information, but the numbers differ. The line representing 25 cases for masked counties (scale on the left) corresponds to 14 cases for unmasked counties (scale on the right). This 11 point difference is consistent throughout the scale. Without careful attention, a consumer might conclude that the 7-day rolling averages for masked counties dropped below those of the unmasked counties. This, however, is not the case. Consider the following graph, where the same data is plotted using the same axes.



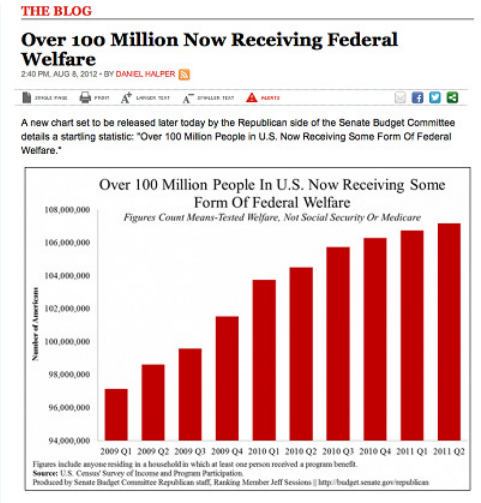
We now see clearly that the masked county 7-day rolling averages stay above those of the unmasked counties. There can be a variety of factors at play in this difference. While the 7-day rolling averages should be scaled to account for differences in population, other related factors could also be at play. Smaller and rural counties were less prone to mask mandates. Larger and more densely populated counties were more likely to have mask mandates. These differences alone could explain the differences in the 7-day rolling averages.

The changes in the 7-day rolling averages could be a better measure for the success of the mandates, but notice how the inclusion of 0 on the vertical scale helps us better gauge the change over time. The variation in the 7-day rolling averages seems less dramatic in the second graph than in the original. The no-mask counties seem to hover around 9 or 10 while the mask counties seem to hover between 16 and19. The first data point makes the change seem quite stark, but a change in a 7-day rolling average indicates that the first measurement was significantly higher than the subsequent days. Would one day of wearing masks have such an immediate effect? Possibly. Such a question warrants further analysis, preferably with each day's raw data.

We have encountered several **bar graphs** in the previous section as we studied relative frequency and frequency distributions. While pie graphs help make comparisons from a part to the whole, bar charts help make comparisons between parts and across different distributions.

? Text Exercise 2.2.3

The 2009 – 2011 data from the U.S. Census Survey of Income and Program Participants was released in USA Today in 2012, and the bar graph below is used. At first glance, the information on the increased number of people on welfare over two years is staggering. What else do you notice about this graph?



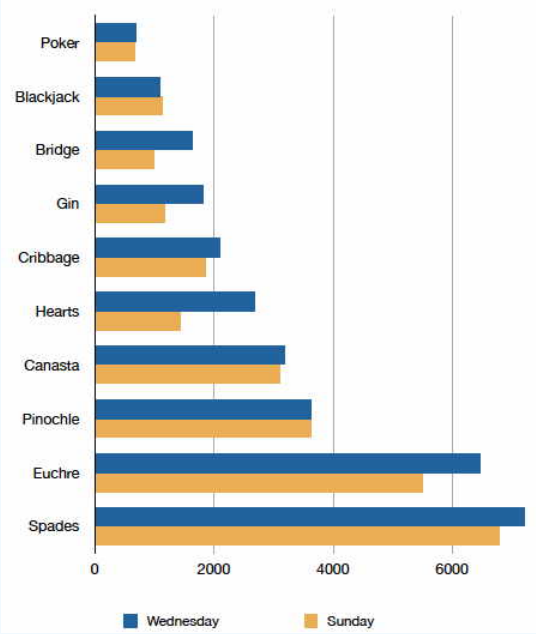
Answer

Although having 108 million people on welfare is not good, we must recognize in this particular graph that the vertical axis begins at 94 million, making the increase look more extreme than it is. Be sure to check the vertical axis to see where it begins. We naturally make mental length comparisons and ratios when looking at bar charts, but recall that ratios are only meaningful if we have a meaningful zero value. When bar charts do not start at a meaningful zero value, like in the graph above, the mental ratio comparisons have no meaning.

We remark here that bar charts displaying ratio data should begin at 0. Unless the reason for beginning at some nonzero value is highlighted and explained, such a chart is more likely to be misleading than informative. We should notice if we are ever shown a bar chart that does not start at 0, as it typically means the differences between the heights are being exaggerated. Even starting at 0 with non-ratio data is misleading, as multiplicative relationships are inherently meaningless. For this reason, we recommend avoiding bar charts for non-ratio data.

? Text Exercise 2.2.4

Consider the graph below that shows the number of people playing card games on the Yahoo website on a Sunday and Wednesday in the Spring of 2001. Here, we have two distributions of game frequencies based on the day. Bar graphs are helpful when comparing distributions taking on the same classes. What can we conclude from such a graph?



Answer

The number of people playing Pinochle was the same on these two days. In contrast, about twice as many people were playing hearts on Wednesday as on Sunday. Blackjack was the only game with more players on Sunday than on Wednesday. Facts like these emerge clearly from a well-designed bar graph. The bars are oriented horizontally rather than vertically. The horizontal format is recommended when you have many categories because there is more room for the category labels.

We can also conclude that there were more players overall on Wednesday than on Sunday. This is because Blackjack was the only game with more players on Sunday than Wednesday, and the margin wasn't large enough to offset the rest. Making conclusions about all the categories collectively is not always easy because bar charts are most conducive to comparing parts to parts.

? Text Exercise 2.2.5

Graphs are pictorial representations of numbers. The following graph, originally from Erickson Times, shows the number of medals per country in the Summer Olympics. Take a minute to observe the graph and see why the pictures in this particular graph may be misleading.

Answer

We naturally expect the representation of the numbers to be proportional to the actual numbers. When looking at Germany's medal count, we see two medals on the graph equal to about 500 medals; we reason that each medal pictured would equal about 250 medals. France only has 523 medals with three units on the graph, and Russia has 999 medals with five units on the graph. The numerical order is correct, but there is no consistent, intuitive correspondence between the units on the graph and the actual number of medals. When using pictures, a graph may look fancy, innovative, or visually appealing but can render the graph misleading and ineffective. The purpose of graphical representations is to disseminate information clearly and quickly. Be cautious.

2.3: Histograms

Learning Objectives

- Distinguish between bar graphs and histograms
- Explore the creation of a grouped frequency distribution and its graphical representation
- Explore how the number of classes affects graphical representations.

Section 2.3 Excel File (contains all of the data sets for this section)

Histograms vs. Bar Graphs

Recall that a histogram is a graph of the distribution of a continuous quantitative variable. Continuity is indicated by eliminating the space between the bars. When the bars have gaps, we have a bar graph representing either a qualitative or discrete quantitative variable. Please note that there are statisticians who distinguish between bar graphs (with gaps) and histograms (without gaps) simply based on the type of variables with bar graphs for qualitative variables and histograms for quantitative variables.

Constructing Histograms

Consider how long it takes to respond to a deer or child suddenly running onto the road while driving a car. We continue to travel at the same speed from the time we see the deer to the time we act (to begin braking, to swerve, etc.); the distance traveled during this time is called the reaction distance. We consider the reaction distance of 642 students in drivers' education when driving at 60 miles per hour. Such a large data set is difficult to understand when presented as a long list of values in a spreadsheet (not provided here). Since the reaction distance is a continuous quantitative variable, we can understand the data better using grouped frequency distributions and histograms.

If we were given that the distances ranged from 46 feet to 167 feet, how might we consider grouping the data? There is no single correct answer. We might choose to go by tens starting at 45 or perhaps starting at 40; we might decide to go by twenties or some other nice number. We recommend constructing several different groupings to see which best represents the data. We show a single example below.

Table 2.3.1: Grouped frequency table for reaction distances

Interval's Lower Limit	Interval's Upper Limit	Class	Class Frequency
39.5	49.5	$(39.5, 49.5]$	3
49.5	59.5	$(49.5, 59.5]$	10
59.5	69.5	$(59.5, 69.5]$	53
69.5	79.5	$(69.5, 79.5]$	107
79.5	89.5	$(79.5, 89.5]$	147
89.5	99.5	$(89.5, 99.5]$	130
99.5	109.5	$(99.5, 109.5]$	78
109.5	119.5	$109.5, 119.5]$	59
119.5	129.5	$(119.5, 129.5]$	36
129.5	139.5	$(129.5, 139.5]$	11
139.5	149.5	$(139.5, 149.5]$	6
149.5	159.5	$(149.5, 159.5]$	1
159.5	169.5	$(159.5, 169.5]$	1

The reaction distances must be broken into mutually exclusive and exhaustive classes, often called **class intervals**. For our example, the first interval is $(39.5, 49.5]$,the second is $(49.5, 59.5]$,etc. Note 59.5 is counted in the second class. The length of a

class interval is called the **class width** and is found by computing the difference between two consecutive lower bounds. The class width is $49.5 - 39.5 = 10$.

The number of reaction distances falling into each interval was counted to obtain the class frequencies. There are 3 reaction distances in the first interval, 10 in the second, etc. We note that class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too "choppy." If this were not the case, we could try a different width.

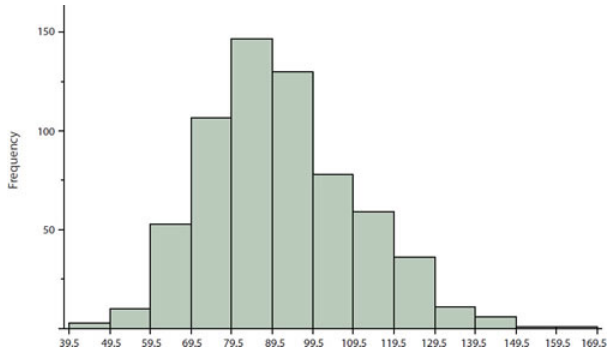


Figure $2.3.1$: Histogram of grouped frequency distribution for reaction distances

In the histogram above, the bar heights represent the frequencies for each of our classes; we could also construct histograms based on relative frequencies. Histograms based on relative frequencies show the proportion of observations in each interval rather than the number of observations. We can change a histogram based on frequencies to one based on relative frequencies by dividing each class frequency by the total number of observations and plotting the quotients on the vertical axis.

Our histogram shows that most reaction distances are in the middle of the distribution, with fewer scores in the extremes. We can also see that the distribution is not quite symmetric: the reaction distances extend to the right farther than they do to the left. The histogram is said to be positively skewed.

To explore these ideas further, we will first utilize Desmos. The following exercise reveals some consequences in changing the number of classes used to construct a histogram.

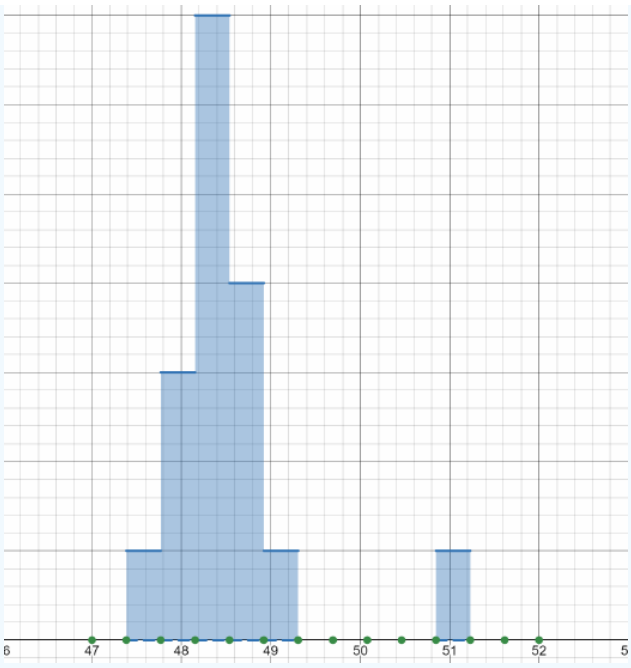
? Text Exercise 2.3.1

Open the Desmos Activity link.

1. Determine what this data represents, find the lower boundary of the first bar of the histogram, determine the class width, and finally describe the shape of this histogram.

Answer

In the far right corner, we can see that the histogram is illustrating times for male swimmers in the 100-meter freestyle. Unfortunately, we do not know when or where these times were collected. By clicking on the points on the x-axis, we can see that the lower bound of our histogram is at 47.385.Notice this does not mean that someone actually swam the 100-meter freestyle at that exact time, but rather, there was one swimmer who swam the 100-meters between $47.385$and 47.769 seconds. The class width can be determined by subtracting consecutive values, as shown in the picture. Each class is 0.385 seconds long. The main portion of this graph is symmetrical from a practical perspective. Given how far and unconnected the last observation is from the rest of the data, it is difficult to say the tail on the right is longer. It seems more likely that far right observation is uncommon.



2. Scroll down to the slider for classes. What happens to the histogram as b gets smaller or larger?

Answer

Notice that having too many classes is essentially looking at each individual piece of data and having too few classes is rather uninformative. Typically, for data sets with fewer than 200 observations, 7 to 10 bins will provide a good representation of the data. With this particular data, however, it seems that 14 classes give us a good view of the data.

We now turn to Excel to familiarize ourselves further with the construction of histograms and with the functionality of Excel.

? Text Exercise 2.3.2

The data set provided for exercise $2.3.2$ in the Section 2.3 Excel file contains the final grade percentages of 260 students.

1. Use the MAX and MIN functions in Excel explained in the provided Excel guides to determine the largest and smallest values in the data set.

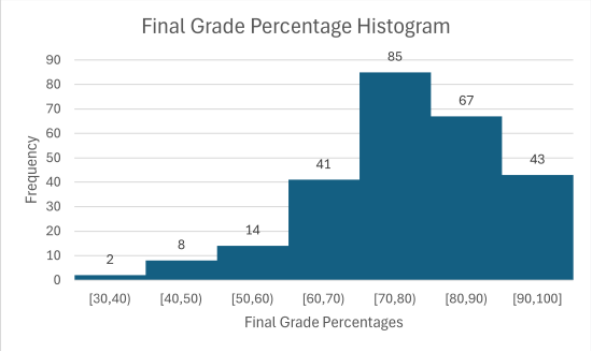
Answer

Using the provided Excel file without altering the columns, the commands $=\mathrm{MAX}(A2:A261)$ and $=\mathrm{MIN}(A2:A261)$ will return the desired values. We thus find that the lowest final grade was 35.73 percent and the highest final grade was 99.76 percent.

2. Knowing that all of our data falls between 35.73 and 99.76, helps ensure that our classes are exhaustive. A natural place to begin constructing a histogram for grade data would be using the typical grading scale for assigning letter grades. As such, the class widths will be 10 with an A being assigned for grades in the interval $[90, 100]$, B for grades in $[80, 90)$, etc. To keep our class widths the same size, continue segmenting the failing grades by 10 as well, rather than just having a class constituting all failing grades. At this point, we caution against the use of Excel's built-in histogram function because, as of 2024, it includes the upper bound of a class in the class. This unfortunate design, however, can be overcome in various ways. See the Excel guide to see how and then construct the histogram. Describe the histogram as symmetric, positively skewed, or negatively skewed. Explain.

Answer

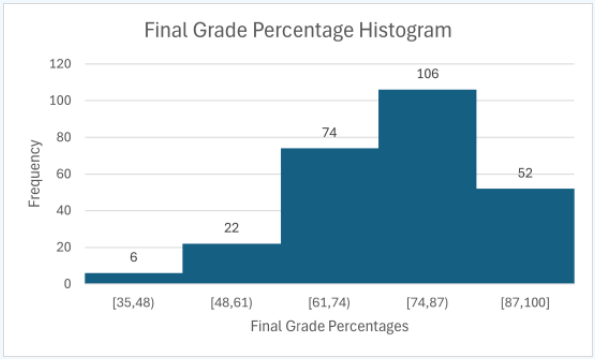
Be sure to label the histogram and various axes with pertinent information. We have included class counts on this histogram for the purposes of checking solutions.

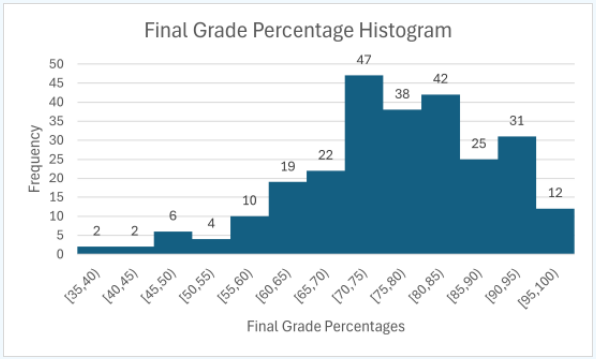


The distribution appears to have a longer tail to the left which would lead to use describing the histogram as negatively skewed.

3. The previous text exercise indicated that a general rule of thumb for constructing data sets with less than 200 observations was to use between 7 and 10 classes. The last histogram consisted of 7 classes and we have over 200 observations but just by 60. Construct two histograms, one with 5 classes and the second with 13 classes. Use 35 as the lower bound of the first class, 100 as the upper bound for the last class, and the same guidelines as the previous histogram in terms of including or excluding the boundaries of the classes. Compare the three histograms.

Answer





Each histogram appears to be negatively skewed. The least pronounced skew is with the histogram constructed from only five classes. When we have 5 or 7 classes, each class increases in frequency until we arrive at the class with the most observations and then each class decreases in frequency as we move beyond. In the histogram with 13 classes, the frequency counts are more volatile going up and down with greater frequency as we progress through the classes.

2.4: Box Plots, Quartiles, and Percentiles

Learning Objectives

- Introduce box plots
- Define quartiles
- Define percentiles
- Calculate percentiles
- Calculate values for a five-number summary

Section 2.4 Excel File (contains all of the data sets for this section)

Using Box Plots to Visualize Data

Frequency distributions and their graphs (bar graphs and histograms) provide insight into data by grouping observations into classes and then determining each class's frequency or relative frequency. The classes depend on the values our data takes on, and there is some freedom regarding the number of classes we might choose to separate our data into.

Another method of graphing ordinal, interval, or ratio level data, called a **box plot** (or a box-and-whisker plot), groups data into four classes based on order, each containing approximately 25% of the observations. Consider the following figure containing three box plots relating students' final grades in statistics across different universities.

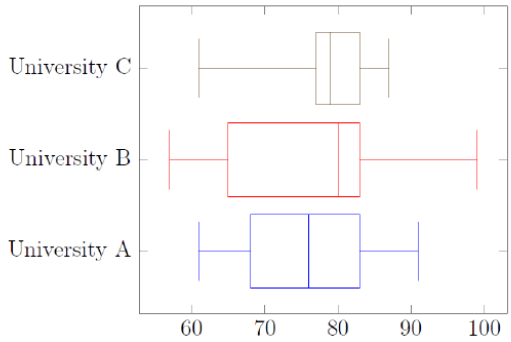


Figure 2.4.1: Box plots of final statistics grades for three universities

We know that there are four classes for each box plot. Note the five vertical lines; these values correspond to the boundaries of the classes. The left-most line corresponds to the data set's smallest value, the minimum. The first class extends from the minimum to the left side of the "box" and includes 25% of the observations; we call the upper bound for this first class, the **first quartile Q_1** (one-quarter of the observations are less than or equal to it). The second class again needs to have 25% of the observations with a lower bound of Q_1 and an upper bound Q_2. We call the upper bound of the second class the **second quartile Q_2**, which is most commonly referred to as the **median,** meaning 50% of the data fall below this value. The third class again needs to have 25% of the observations with a lower bound of Q_2 and an upper bound of the **third quartile Q_3**; 75% of the observations are less than or equal to Q_3. The final class consists of the remaining 25% of observations with lower bound Q_3 and upper bound of the largest value, the **maximum**, of the data set. The first and fourth classes form the whiskers of the box plot, while the second and third form the box. These five numbers (**minimum**, **Q_1, Q_2, Q_3,** and **maximum)** form what we call the **five-number summary** of the data.

Text Exercise 2.4.1

Consider Figure 2.4.1 above and classify each box plot as positively skewed, negatively skewed, or symmetric. Explain.

Answer

We preface our answer by noting that box plots, like histograms and bar graphs from grouped frequency distributions, are formed by grouping observations. Since the graph does not provide information about each data value, we are primarily making a claim about a characteristic of the graphical representation.

Recall that a graph is positively skewed if the right tail extends further than the left tail and negatively skewed if the left tail extends further than the right tail. A graph is symmetric if we can fold it in half so that the left and right sides roughly match. The tails and the whiskers fall in similar parts of the graphs discussed. If one whisker is longer, we can say that the box plot is skewed in the direction of the longer whisker. We classify the box plot of University C as negatively skewed and the box plot of University B as positively skewed. The whiskers seem to be of equal length for University A, but that is not enough to assert symmetry. We also want the two halves of the box to be the same. This is the case for the box plot of University A. We classify the box plot of University A as symmetric.

Given a data set, we can quickly identify the minimum and maximum values. Determining the quartiles, however, presents more of a challenge. With an ordered data set, we understand where a quarter, half, and three-quarters of the data would fall. Consider the following data sets with their five-number summaries.

Observations	2	4	6	8	10	12	14	16
Five Number Summary	min		Q_1		Q_2		Q_3	max

Observations	2	4	4	8	10	12	14	16
Five Number Summary	min		Q_1		Q_2		Q_3	max

Observations	2	4	6	8	10	12	14	16	18
Five Number Summary	min		Q_1		Q_2		Q_3		max

Figure 2.4.2: Three data sets with intuitive placement of the box plot boundaries

We need help determining our five-number summary for these data sets.

The first two data sets each contain 8 observations, while the third contains 9 observations. We can easily group the first two data sets in groups of 2 to get 25% of the observations in each class. With 9 observations, however, we cannot get exactly 25% in each class. However, note that there are equal numbers of observations above and below Q_2 for each data set.

If we try to attach values to quartiles, we face another challenge. Let us begin with Q_1. In the first data set, we see that Q_1 naturally falls between 4 and 6, but what value should we assign? Our box plots would look significantly different if we used 6 as opposed to 4. There is no easy solution to this challenge, and statisticians have developed a variety of approaches. We will provide a simple approach later in this section; please remember that it is not the only approach. However, most of the various approaches produce measures that are reasonably close to each other.

We face another challenge when we try to understand Q_1 in the second data set. We would naturally assign a value of 4 to Q_1 because the only number between 4 and 4 is 4. We wanted 25% of the observations to be less than or equal to Q_1, but we have 3 values that are less than or equal to 4 in our data set, that is 37.5% of our observations.

We highlight these challenges to frame our understanding appropriately. We use box plots, quartiles, and percentiles (which we will get to shortly) to get a general, intuitive feel about our data using methods that may differ from field to field, statistician to statistician, and program to program. When consuming statistics or conducting analysis, know which method is in use.

Quartiles are descriptive statistics that express at what values there will be about 25%, 50%, or 75% of the observations at or below that value. There is nothing extraordinarily unique about 25%, 50%, or 75%. We could choose 10% or 99%. When we expand our ideas to include different percentages of observations, we call them **percentiles**. Q_1 is the 25^{th} percentile. Q_2 is the 50^{th} percentile. Q_3 is the 75^{th} percentile.

Percentiles have utility beyond building summary visualizations; they help us understand how individual observations compare to the entire data set. They measure relative position within an ordered data set. For example, a test score by itself is usually difficult to interpret. For instance, if one of us had a score on a measure of shyness of 35 out of a possible 50, we would have little idea how shy that person was compared to others. It would be helpful to know the percentage of people with equal or lower shyness scores. If 65% of the scores were at or below this person's score, then the score would be at the 65^{th} percentile.

Text Exercise 2.4.2

1. If Helen's score was at the 95^{th} percentile, what percentage of scores are at or below Helen's?

Answer

The percentile means that 95% of the scores are at or below Helen's score.

2. If the scores ranged from 1 to 100 on an exam and Helen earned a score of 95, does this necessarily mean that her score is at the 95^{th} percentile?

Answer

No, the percentile gives a relative position of the scores. The number of scores at or below her score determines the percentile measure. If everyone did well and only 70 of the scores fell at or below Helen's, she would be at the 70^{th} percentile even though she got 95 out of a 100 points.

Calculating Percentiles

We already indicated that there are several different ways to calculate quartiles. This is because quartiles are percentiles, and there are several ways to calculate percentiles, which may lead to different values in different situations. The method that we present is one of the simplest calculations.

The P^{th} percentile is a value such that $P\%$ of the observations fall at or below that value. We need the data to be counted and ordered from smallest to largest. Let n be the number of observations in our data set. Next, we calculate the number of observations that make up $P\%$ of the observations. We call this number the rank R of the percentile.

$$R = \frac{P}{100} \cdot n$$

Now there are two possibilities for the R value; either it will be a natural number $\{1, 2, 3, 4, 5, \dots\}$ or not.

- If R is a natural number, we find the midpoint between the R^{th} and $(R+1)^{st}$ values.
- If R is not a natural number, round R up to the following natural number and take the value in that position.

Note: Classifications of Numbers

The real number system has the following designations.

Natural Numbers: $1, 2, 3, \dots$
Whole Numbers: $0, 1, 2, 3, \dots$
Integers: $\dots, -3, -2, -1, 0, 1, 2, 3, \dots$
Rational Numbers: Any number that can be written as a fraction
Irrational Numbers: Any number that cannot be written as a fraction. Examples include e, π, $\sqrt{2}$

Text Exercise 2.4.3

Consider the 20 quiz scores shown in the table below and compute the five-number summary and 82^{nd} percentile. After completing the calculation by hand, use the Section 2.4 Excel file to calculate each percentile using the functions PERCENTILE.INC and PERCENTILE.EXC. Compare the values.

Table 2.4.1: 20 quiz scores with corresponding rank

Number	**4**	**4**	**4**	**5**	**6**	**6**	**6**	**6**	**6**	**6**	**7**	**7**	**8**	**8**	**8**	**9**	**9**	10	10	10
Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	10	11	12	13	14	15	16	17	18	19	20

Answer

We first note that the data is already ordered from smallest to largest with 20 observations. A secondary row has been created to index the observations. Note that the row heading is Rank; consider how this ties back to rank R in our calculation.

Let us begin with the five-number summary.

The minimum value is 4.

Q_1, the 25^{th} percentile. $R = \frac{25}{100} \cdot 20 = \frac{1}{4} \cdot 20 = \frac{20}{4} = 5$. Note 5 is a natural number. We then look at the 5^{th} and 6^{th} observation values, which are both 6 and find the midpoint between them. Thus $Q_1 = 6$. Using Excel, we get 6; they are all the same.

Q_2, the 50^{th} percentile. $R = \frac{50}{100} \cdot 20 = \frac{1}{2} \cdot 20 = \frac{20}{2} = 10$. Note 10 is a natural number. We then look at the 10^{th} and 11^{th} observation values, which are 6 and 7, respectively and find the midpoint between them. Thus $Q_2 = \frac{1}{2}(6+7) = \frac{13}{2} = 6.5$. Excel computes 6.5 and 6.5;they are all the same.

Q_3, the 75^{th} percentile. $R = \frac{75}{100} \cdot 20 = \frac{3}{4} \cdot 20 = \frac{60}{4} = 15$. Note 15 is a natural number. We then look at the 15^{th} and 16^{th} observation values, which are 8 and 9, respectively and find the midpoint between them. Thus $Q_3 = \frac{1}{2}(8+9) = \frac{17}{2} = 8.5$. Excel gives 8.25 and 8.75;they are all different.

The maximum value is 10.

Let us look at the 82^{nd} percentile. $R = \frac{82}{100} \cdot 20 = \frac{41}{50} \cdot 20 = \frac{820}{50} = \frac{82}{5} = 16.4$. Notice 16.4 is not a natural number. We must look at the 17^{th} observation value showing the 82^{nd} percentile is 9. Excel calculates 9.22 and 9; they are different and the same respectively.

We have seen that different methods of calculation can produce slightly different values. With large data sets, we generally resort to technology to produce our measures and might not have control over the precise methodology employed therein. As such, we remember that percentiles provide rough measures for the distribution of our data sets and nuance our understanding that roughly this percent of observations fall below roughly this value. When large data sets or limited time make hand computation prohibitive, we recommend using functions such as PERCENTILE.INC and QUARTILE.INC.

Consider the preceding example. If we looked at the 83^{rd} percentile $(R = 16.6)$ which would also be the 17^{th} value, which is 9. So, the 83^{rd} percentile is the same as the 82^{nd} percentile. This happens since there are only 20 data points; we cannot subdivide 20 indefinitely. The two percentiles would typically differ with large data sets.

Box Plots: Constructing and Interpreting

? Text Exercise 2.4.4

As part of the "Stroop Interference Case Study," students in introductory statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. Compare the scores for the 15 men and 30 women who participated in the experiment by making separate box plots for each gender.

Table 2.4.2: Women's times (left) and men's times (right)

14	17	18	19	20	21		16	19	23
15	17	18	19	20	22		17	20	24
16	17	18	19	20	23		18	22	26
16	17	18	20	20	24		19	22	26
17	18	18	20	21	24		19	23	28

Answer

To construct box plots, we need the five-number summaries.

Table 2.4.3 Five number summaries for the data presented in Table 2.4.1

	Females	Males	Box Plot Component
Minimum	14	16	End of Left Whisker
Q_1	17 ($R = \frac{25}{100} \cdot 30 = 7.5$)	19 ($R = \frac{25}{100} \cdot 15 = 3.75$)	Left Side of Box
Q_2=median	18.5 ($R = \frac{50}{100} \cdot 30 = 15$)	22 ($R = \frac{50}{100} \cdot 15 = 7.5$)	Line in Box
Q_3	20 ($R = \frac{75}{100} \cdot 30 = 22.5$)	24 ($R = \frac{75}{100} \cdot 15 = 11.25$)	Right Side of Box
Maximum	24	28	End of Right Whisker

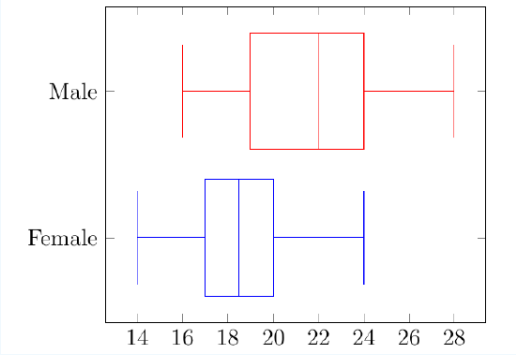


Figure 2.4.3 Box plots for male and female times for naming the colors of various rectangles

The men tended to take longer than the women. About 25% of male times were longer than the maximum female time. At least 75% of the male times were longer than the median female time.

Text Exercise 2.4.5

Suppose data came from a task that aims to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. The time to reach the target was recorded on each trial. The box plots of the two distributions are shown below. What can we conclude by looking at the two box plots?

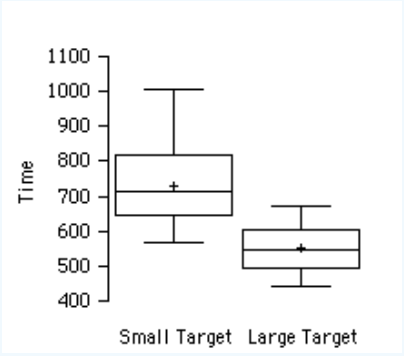


Figure 2.4.4: Box plots for the response times by small and large target

Answer

We can see that although there is some overlap in times, it generally took longer to move the mouse to the small target than to the large one. The minimum time for the small target is longer than the median time of the large target. At least 50% of times for the small target are longer than all of the large target times.

Text Exercise 2.4.6

Construct two data sets, treated as observations from a discrete quantitative variable, consisting of 12 values each so that the box plots are identical, but the bar graph of one data set is perfectly symmetric while the other is not.

Answer

If the box plots are going to be identical, the five-number summaries must be the same. When arranged from least to greatest, Q_1 is the average of the 3^{rd} and 4^{th} values, Q_2 is the average of the 6^{th} and 7^{th} values, and Q_3 is the average of the 9^{th} and 10^{th} values. We can construct a perfectly symmetric data set by pairing observations by proximity to the center. Since the 6^{th} and 7^{th} values are in the middle, they would be paired together, the 5^{th} with the 8^{th}, and the 4^{th} and 9^{th}, and so forth. We want the values in these positions to be equally distant from the median value.

We can start by picking the first data set that is not perfectly symmetric. We will produce such a data set if we use a pattern of one observation value followed by a different value repeated twice. We picked even numbers starting at 0 to ensure that our quartiles were nice values. Not all procedures would produce a data set suitable for this example because we need our second data set to be symmetric. We must check this because we want the minimum and maximum to be equally far from the median and the first and third quartiles.

$$\{0, 2, 2, 4, 6, 6, 8, 10, 10, 12, 14, 14\}$$

We have fixed our five number summary: $\min = 0$,$Q_1 = 3$, $Q_2 = 7$, $Q_3 = 11$, and $\max = 14$. Our procedure produced values allowing us to produce a symmetric data set with the same five-number summary. The average of two of the same numbers is that number, so, for ease, we can start our symmetric data set with the following numbers.

$$\{0, \ , 3, 3, \ , 7, 7, \ , 11, 11, \ , 14\}$$

All we need to do is fill in the remaining spots, ensuring symmetry and preserving order.

$$\{0, 2, 2, 4, 6, 6, 8, 10, 10, 12, 14, 14\}$$
$$\{0, 1, 3, 3, 4, 7, 7, 10, 11, 11, 13, 14\}$$

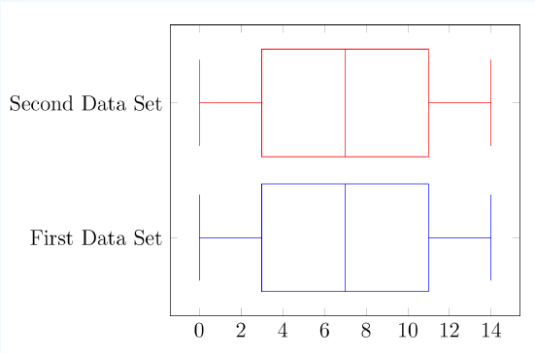


Figure 2.4.5 Two identical box plots

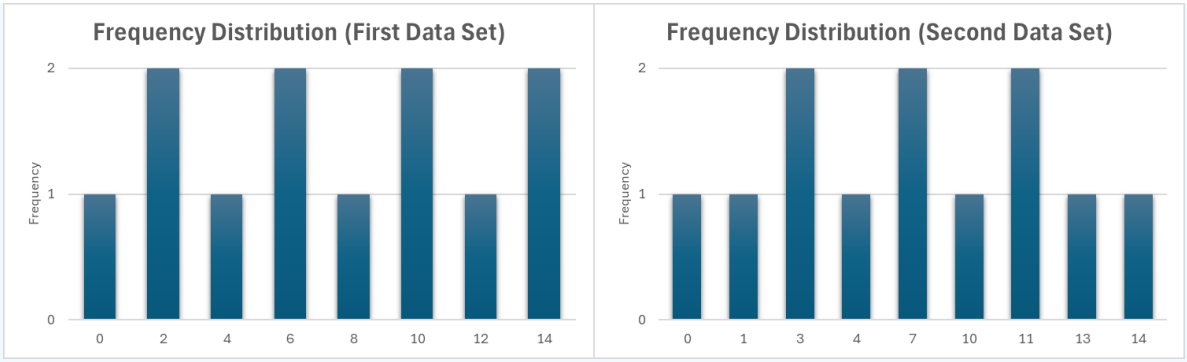


Figure 2.4.6 Non-identical bar graphs of the two data sets that produced identical box plots.

2.5: Measures of Central Tendency

Learning Objectives

- Discuss common measures of central tendency: mean, median, and mode
- Introduce the trimmed mean

Section 2.5 Excel File (contains all of the data sets for this section)

Introduction to Measures of Central Tendency

We understand data by looking at distributions (graphs and tables) and box-and-whisker plots. With relative frequency distributions, we determined classes and then computed the percentage of observations in each class. With box-and-whisker plots, we determined four classes each with 25% of the observations. These were constructed using descriptive statistics and allowed us to see where and how the data values fell; we could see the distribution of the data. In this section, we discuss descriptive statistics that indicate the center of the data. There are many different ways to define the center of a data set; each measure has strengths and weaknesses. We discuss the three most common measures of central tendency: mode, median, and mean.

Mode

When examining a frequency distribution, either a table or a graph, our attention often gravitates to the highest frequency: the value that occurs the most. Sometimes, this highest frequency occurs in multiple classes. We call the class(es) with the highest frequency the **mode(s)**. If there is only one class with the highest frequency, we call the distribution **unimodal;** otherwise, we call it **multimodal**. The mode is a measure of central tendency by describing the classes that occur most frequently; the distribution is often centered around these most common values. The mode can be computed for any variable regardless of its level of measurement. It is the only measure we will discuss that is defined for nominal data.

Text Exercise 2.5.1

Recall the frequency distribution of colors of candies in a bag of M&M's from our previous discussion.

Table 2.5.1: Frequencies and Relative Frequencies of Sampled M&M's

Color	Frequency	Relative Frequency
Brown	17	$\frac{17}{55} \approx 0.309$
Red	18	$\frac{18}{55} \approx 0.327$
Yellow	7	$\frac{7}{55} \approx 0.127$
Green	7	$\frac{7}{55} \approx 0.127$
Blue	2	$\frac{2}{55} \approx 0.036$
Orange	4	$\frac{4}{55} \approx 0.073$

1. Show that this data set is unimodal and give the mode.

Answer

The mode is the value (characteristic) that appears most frequently. Red is the only mode since red appears 18 times, and all other colors appear fewer than that. Note that yellow and green are not modes even though 7 appears twice.

2. Show the set of colors could be multimodal after consuming one candy.

Answer

We would have to reduce the number of red candies to make the set multimodal. If we eat one red candy, we would have 17 brown and 17 red candies. Since every other color appears fewer than 17 times, we have two modes: red and brown. The set would be multimodal, specifically **bimodal** (since there are two modes).

3. What is the minimum number of candies one would have to eat for orange to be the only mode?

Answer

Since there are only 4 orange candies, we would have to reduce the number of every other color to less than 4. If we ate 14 brown, 15 red, 4 yellow, and 4 green candies, we would have 3 of each of those colors, 2 blues, and 4 orange making orange the mode. We need to eat, at minimum, $14+15+4+4=37$ candies. YUM!

It is easy to see the possible issues with the mode as a measure of central tendency. Consider the following data set: $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 1000, 1000$. The mode is 1000, but is that a reasonable value for the center of this data? The use of the mode is often minimal with quantitative data.

Median

When constructing a box-and-whisker plot, we computed five measures: minimum, first quartile, second quartile, third quartile, and maximum. Each of these measures relative position and relies on ordering the data. A natural measure of central tendency would be a value that splits the data evenly below and above, such as the second quartile, the 50^{th} percentile. We generally refer to it as the **median**, one of the most common measures of central tendency. Since the median requires an ordering from smallest to greatest, it cannot be computed for nominal variables, but it can be calculated for ordinal, interval, and ratio variables.

Arithmetic Mean

The third measure, called the arithmetic mean, is arguably the most common measure of central tendency. Bar graphs may remind us of geometric figures placed along a scale, making us wonder: what is the center of mass? This would be the point at which the figure would balance if it were propped up only at that point. How might we find such a point? Taking each observation as equally important, we assign each observation an equal weight w distributed evenly across a uniform-sized block and stack the blocks on our scale with the blocks centered on their value. Consider Figure 2.5.1.

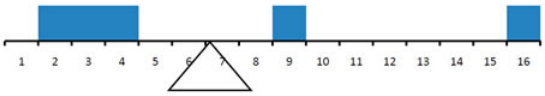


Figure 2.5.1: A simple distribution balanced upon its center of mass

There is only one location c that our distribution balances. When it is balanced, there is no motion. The torque from the blocks to the right of c, eliciting a clockwise motion, is equal to the torque from the blocks to the left of c, eliciting a counterclockwise motion. The total torque is equal to the sum of the torques from the individual blocks, and the torque from each block is equal to the distance the block is from c multiplied by the weight of the block, yielding the following equation:

$$(c-2)w+(c-3)w+(c-4)w=(9-c)w+(16-c)w.$$

Note that every term has the same weight, w, which cancels algebraically. Move all of the c terms together and constants together to produce the equivalent equation

$$5c=2+3+4+9+16=34$$

$$c=\frac{34}{5}$$

We could have chosen to factor out a negative sign from the left side to yield:

$$-((2-c)+(3-c)+(4-c))=(9-c)+(16-c)$$

Meaning that our equation is equivalent to

$$0=(2-c)+(3-c)+(4-c)+(9-c)+(16-c)$$

We may generalize to other data. Let x_i represent our i^{th} value from our collection of n data values. To find our center of mass c, we need the torques to balance, which is equivalent to the following:

$$0 = \sum_{i=1}^{n}(x_i - c) = x_1 - c + x_2 - c + \ldots x_n - c = \left(\sum_{i=1}^{n} x_i\right) - n \cdot c$$

Meaning

$$c = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This is the familiar formula for the arithmetic mean, what many call "average." The **arithmetic mean** of a data set is the center of mass for its frequency distribution and is a measure of central tendency.

The arithmetic mean plays a significant role in statistics and throughout this course. There are several different types of means, but given the prevalence of the arithmetic mean, we will refer to the arithmetic mean simply as the mean. We have standard notation to differentiate the mean as a statistic, denoted $\bar{x}$, from the mean as a parameter, denoted μ. This latter symbol is the lowercase Greek letter **mu**. Recall that parameters are generally denoted with Greek letters and refer to properties of a population, not a sample. Notice the similarity in the formulas for computations below:

$$\bar{x} = \frac{\sum x_i}{n} \qquad \mu = \frac{\sum x_i}{N}$$

Note: the summations used in these formulas do not include any indexing information. When this is the case, sum over all observations.

Since the arithmetic mean requires that the differences between values have meaning, it cannot be computed for nominal or ordinal variables but can be computed for interval and ratio-level data.

? Text Exercise 2.5.2

Consider the following distributions. Determine which of the common measures of central tendency are indicated by the blue and pink bars below the scaling axis. Explain your reasoning.

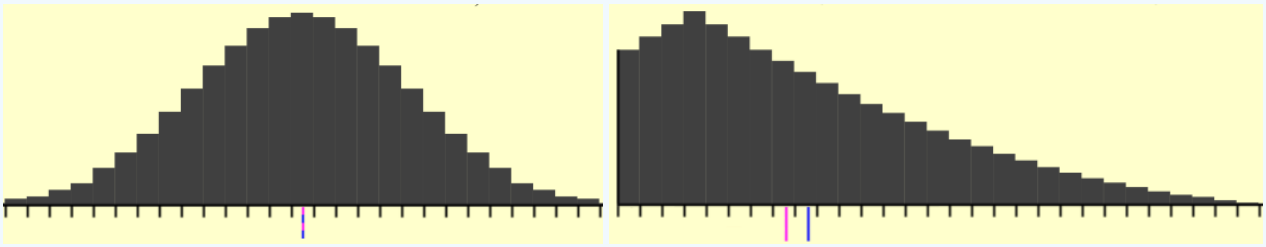


Figure 2.5.2: A symmetric distribution (left) and a positively skewed distribution (right)

Answer

There are three standard measures of central tendency: mode, median, and mean. The distribution on the left is unimodal and symmetric. The distribution on the right is also unimodal but is positively skewed. The mean, median, and mode are all equal in the left distribution. This follows from the fact that it is unimodal and symmetric. (Can you explain why?) The mode of the right distribution is less than both colored bars on the scale axis; thus, the bars do not represent the mode.

Determining which bar represents the mean is a little more complicated. Recall that the number of observations above the median and below the median should be equal. The median splits the number of observations in half, but it is difficult to tell which bar has the same number of observations on both sides. When analyzing relationships, it is best to alter only one variable at a time. If we increase the maximum value (move it to the right), the median won't change. However, the mean would become larger to keep the figure balanced. A positive skew moves the mean to the right, and the mean is larger than the median. The blue bar is the mean, and the pink bar is the median.

In the previous exercise, we decided that the mean would be a larger value than the median in a positively skewed data set. A similar argument yields that the mean takes on smaller values than the median in negatively skewed data. This is because the mean incorporates every value into its computation, while the median only cares about the relative position of the values. We recommend computing the median as the better measure of center when the data is skewed or has values of a more extreme magnitude than the rest.

? Text Exercise 2.5.3

Determine which measures of central tendency are appropriate. Which would be the most appropriate? Explain.

1. Evaluations rated on a scale of 1 through 5

Answer

Rating scales are measured on the ordinal scale. While statisticians debate the legitimacy of averaging ordinal data, we recommend avoiding the practice. Median and mode are eligible candidates. The median is used more frequently than the mode; it is our measure of choice.

2. Salaries

Answer

Salaries are measured on the ratio scale making mode, median, and mean eligible candidates. However, salary data is often positively skewed, making the median a better choice.

3. Heights

Answer

Heights are measured on the ratio scale. Again, the mode, median, and mean are eligible candidates. Heights are not generally highly skewed, making the mean a better choice.

4. Shirt sizes

Answer

Shirt sizes are measured on the ordinal scale. Between mode and median, we would choose the median

5. Favorite candies

Answer

Favorite candies are measured on the nominal scale. The only option is mode.

? Text Exercise 2.5.4

Create a data set consisting of 10 observations such that the mean is 15, the median is 14, and the mode is 17.

Answer

Use the Section 2.5 Excel file to check your solution by typing your 10 values in the second column. The cells with the running calculations will turn green when that aspect of the data set is correct. There are infinitely many possible data sets. Do not fret if your solution is different than your classmates' solutions.

Sometimes, working backward can be difficult. First, consider what each measure says about the data. If the mean is 15 with 10 observations, the sum of all the values needs to be 150. If the median is 14, the average of the 5^{th} and 6^{th} values in rank order, must be 14. If the mode is 17, 17 must appear the most number of times (at least twice, if every other number only appears once). Start with the most restrictive measures and work through them all. We recommend starting with the median, then the mode, and ending with the mean.

Trimmed Mean

When data is skewed or has values more extreme than the rest, we recommended using the median as a measure of central tendency. There is a less common measure, a hybrid between the mean and median, that can also be used. It is called the trimmed mean; its definition differs across the literature, but its underlying idea is consistent. When we have data of this type, the observations that significantly affect the mean value are the extreme values of the data. To mitigate their influence, we trim a certain percentage of the observations from both the top and the bottom and compute the mean on the remaining data.

When running across the trimmed mean in literature or research, check what definition is being used, as there are subtle differences that are good to be aware of. We now provide our working definition of the $p\%$-trimmed mean. Trim $p\%$ of the data from the top and $p\%$ of the data from the bottom for a total of $2p\%$ and then compute the mean of the remaining data. If $p\%$ is not a whole number, remove the smallest number of observations such that at least $p\%$ of the observations are removed.

? Text Exercise 2.5.5

Compute the mode, median, mean, and 10%-trimmed mean for the following sample data.

$\{10, 6, 5, 25, 10, 11, 17, 13, 15, 13, 19, 10\}$

Answer

Ordering and counting the data for the median and 10%-trimmed mean is necessary.

$\{5, 6, 10, 10, 10, 11, 13, 13, 15, 17, 19, 25\}$ with $n = 12$

Mode: 10 occurs most frequently with three occurrences. 13 comes in at second with two occurrences. The mode is 10.

Median: We must first compute our rank $R = \frac{50}{100}(12) = 6$. Since it is a natural number, we look at the 6^{th} and 7^{th} values, 11 and 13 respectively. Our median is the midpoint, $\frac{1}{2}(11 + 13) = 12$

Mean: $\bar{x} = \frac{5+6+10+10+10+11+13+13+15+17+19+25}{12} = \frac{154}{12} = 12\frac{5}{6} = 12.8\overline{3}$

10%-trimmed mean: We must first compute how many observations we must remove. 10% of 12 is 1.2. This is not a whole number; we, therefore, remove two observations from the bottom and two from the top. We find the average of the following set:

$\{10, 10, 10, 11, 13, 13, 15, 17\}$

10%-trimmed mean$= \frac{10+10+10+11+13+13+15+17}{8} = \frac{99}{8} = 12\frac{3}{8} = 12.375$

Notice we get an incorrect value for the trimmed mean unless we first sort the data.

2.6: Measures of Dispersion

Learning Objectives

- Explore several measures of dispersion in data
- Develop measures of dispersion:
  - Range
  - Interquartile Range
  - Mean Absolute Deviation
  - Variance
  - Standard Deviation
- Compute various measures of dispersion

Section 2.6 Excel File (contains all of the data sets for this section)

What is Dispersion?

Consider the two histograms in Figure 2.6.1 representing scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, we can see that the distributions are quite different. The scores on Quiz 1 are more densely packed than the scores on Quiz 2. The differences of scores were much greater on Quiz 2 than on Quiz 1.

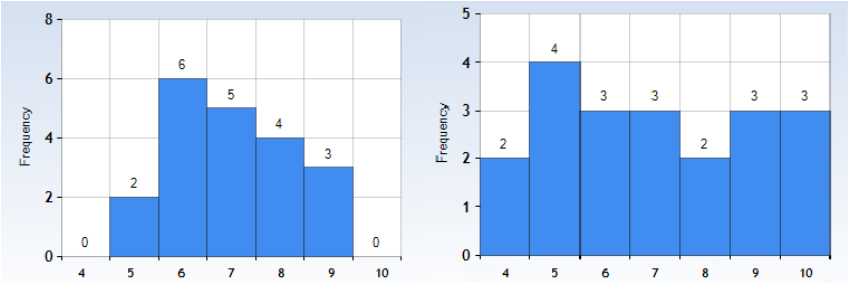


Figure 2.6.1: Histograms for Quiz 1 (left) and Quiz 2 (right)

The terms variability, spread, and dispersion are synonyms. They refer to how varied data are in a data set or how spread out the distribution of the data is. In this section we will discuss measures of the dispersion of a distribution. We seek a single number to describe how spread out or dispersed the data is. There are many ways to measure "dispersion," and no single measure gives complete insight into the data's dispersion. We will examine five frequently used measures of dispersion: the range, interquartile range, mean absolute deviation, variance, and standard deviation.

Range

The range is the most straightforward measure of dispersion to calculate. As a warning, the term "range" is used in multiple ways, so do not confuse the statistical use of this word with other uses, such as in algebra. Recall that our summary measures tend to be given as a single value so, in statistics, the **range** is simply the highest data value minus the lowest data value, that is:

$$\text{range} = \text{maximum} - \text{minimum}.$$

Since we are subtracting data values, we must work with interval or ratio-level data for the range to have meaning; we do not have a range measure in nominal or ordinal level data.

Text Exercise 2.6.1

Determine the range of the following group of numbers.

$$10, 2, 5, 6, 7, 3, 4$$

Answer

The highest number is 10, and the lowest number is 2, so $10 - 2 = 8$. The range is 8. These values are within 8 units from each other.

? Text Exercise 2.6.2

Now consider the two quizzes in Figure 2.6.1. What is the range of each quiz?

Answer

On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4.

The range on Quiz 2 is larger: the lowest score is 4 and the highest score is 10. Therefore, the range is 6.

Since Quiz 1 has a smaller range, we can say that Quiz 2 is more spread out than Quiz 1.

The range is a quick way to get a rough idea of the spread of the data. However, it is a very coarse measure since it depends on only two data points. The sets $\{0,5,5,5,5,10\}$ and $\{0,0,0,10,10,10\}$ have the same range but would not be called equally dispersed. We must investigate other measures.

Interquartile Range

A similar measure to the range is the **interquartile range (IQR)**. The IQR is the range of the middle 50% of the scores in a distribution. We can understand this visually as the length of the box in the box plot.

$$\text{IQR} = 75^{th} \text{ percentile} - 25^{th} \text{ percentile} = Q_3 - Q_1$$

? Text Exercise 2.6.3

In Section 2.4, we looked at data from men and women on the Stroop Test. In the women's data, the 25^{th} percentile is 17, the 50^{th} percentile is 18.5, and the 75^{th} percentile is 20. For the men, the 25^{th} percentile is 19, the 50^{th} percentile is 22, and the 75^{th} percentile is 24. Calculate the IQR for men and women.

Answer

Women: $\text{IQR} = 20 - 17 = 3$

Men: $\text{IQR} = 24 - 19 = 5$

Measures of dispersion relay how spread out the data is. The range of a data set gives the distance between the minimum and maximum values. Similarly, the IQR of a data set gives a distance, but this time the distance is the smallest length of an interval such that the central 50% of observations could fall into the interval. Larger values of range and IQR indicate that the data set is more spread out. One can have data with a large range and small (\text{IQR},\) indicating large dispersion in the entire data set, and yet the central 50% of the data is not varied in comparison.

As we mentioned earlier, these measures do not provide a complete understanding of the data set since they depend on only a few values. As we progress, we will discuss measures that incorporate all data values into the calculation, but even these measures of variability will not provide a complete understanding: better, yes; complete, no.

? Text Exercise 2.6.4

For each measure of dispersion discussed so far, range and IQR, construct two data sets (each with 10 values) that have the same measure of dispersion value but starkly different degrees of spread when looking at the data.

Answer

It is good to remember that there are many different solutions to these questions.

Let us begin with range. If our two data sets are to have the same range, the difference between maximum and minimum values must be the same. We can pick that value freely; let us say 20. We could have both minimum and maximum values be the same, or we could have them be different. Let us keep them the same.

$$\{10, \ , \ , \ , \ , \ , \ , \ , \ , 30\}$$
$$\{10, \ , \ , \ , \ , \ , \ , \ , \ , 30\}$$

Here again, we had the freedom to pick at least one of our values. Once 30 was chosen as a maximum, 10 was forced to be our minimum.

We now need to think about how we could have different degrees of spread. We could have values spread fairly evenly from 10 to 30 in one data set and have them closely packed around one value in the other.

$$\{10, 13, 15, 17, 19, 21, 23, 25, 27, 30\}$$
$$\{10, 20, 20, 20, 20, 20, 20, 20, 20, 30\}$$

Now let us look at IQR. If our two data sets are to have the same IQR, the difference $Q_3 - Q_1$ needs to be the same. As with range, we could have Q_1 and Q_3 be the same or different. Since we have one example where we had the same values, we will make them different. Let us again choose 20 for our value. Since $n = 10$, Q_1 is the third value in our ordered data set, and Q_3 is the eighth value in our ordered data set.

$$\{\ ,\ , 10,\ ,\ ,\ ,\ , 30,\ ,\ \}$$
$$\{\ ,\ , 15,\ ,\ ,\ ,\ , 35,\ ,\ \}$$

What sort of differences could we have in our data sets to elicit different degrees of spread? We already considered a reasonably uniform spread and one centered on a single value. We could have a greater spread outside than inside our box as opposed to having two clusters centered at Q_1 and Q_3.

$$\{-50, -40, 10, 14, 20, 22, 26, 30, 80, 90\}$$
$$\{13, 14, 15, 16, 16, 34, 34, 35, 36, 37\}$$

Deviation

Measures of dispersion give us an idea about how spread apart our data are. Our previous measures incorporated only some of the data values. One way to include all of the data is to compare how far away each piece of data, call it x, is from some specific value v. We call the difference $x - v$ the **deviation from v.** A data value's distance from v, which would be $|x - v|$, is called the **absolute deviation from v.** There are many possible options for v. The most common choice of v is the mean. Once we have all the deviations, we must decide what to do with them since a measure of dispersion is a single value. Two options are summing up and averaging the deviations.

Consider the distribution of the five numbers 2,3,4,9,17.

Table 2.6.2: An example of various deviations

Values	Deviations from the mean	Deviations from the median	Absolute deviations from the mean	Absolute deviations from the median
2	$2-7=-5$	$2-4=-2$	$\|2-7\|=5$	$\|2-4\|=2$
3	$3-7=-4$	$3-4=-1$	$\|3-7\|=4$	$\|3-4\|=1$
4	$4-7=-3$	$4-4=0$	$\|4-7\|=3$	$\|4-4\|=0$
9	$9-7=2$	$9-4=5$	$\|9-7\|=2$	$\|9-4\|=5$
17	$17-7=10$	$17-4=13$	$\|17-7\|=10$	$\|17-4\|=13$
Sum	0	15	24	21
Average Deviation	0	3	$\frac{24}{5}=4.8$	$\frac{21}{5}=4.2$

Perhaps the final value of 0 in the "Deviations from the mean" column surprised some of us. After some thought, it should make sense that some values are above the mean and others are below the mean.

Examining the table can help us gain a deeper understanding. The last two columns look at the absolute values of the deviations rather than the deviations themselves. Deviation is similar to displacement, and the absolute value of deviation is similar to distance. When we sum our deviations, values below our central value contribute negatively, while values above our central value contribute positively and cancel each other out.

Note: Displacement and Distance

Displacement is the difference between the initial position and the final position.

Distance is the path length from our initial position to our final destination.

We are summing up all of the deviations from the mean. If we put this in summation notation, we arrive at

$$\sum (x - \bar{x})$$

Hopefully, we recognize this from our discussion on central tendency. The mean as the balance point is the value that makes this sum equal 0; this will be the case regardless of the data set. Thus, the average deviation from the mean is always 0.

We want to avoid the cancellation with summing deviations; there is more than 0 spread in the data set. The absolute value of the deviations finds the distance each observation is from the central value. The sum of the absolute deviations can be considered the total distance our observations are from our central value. Since the total distance is affected by the number of observations, we prefer to use the **mean absolute deviation from v (MAD).** We can understand the MAD as the average distance the various data values are from the central value. Again, larger MAD values indicate that the data is spread to a greater degree.

$$\text{MAD} = \frac{1}{n}\sum |x_i - v|$$

We will not prove this, but if v is chosen to be the median, the MAD is minimized. Try to convince yourself that this is true using Excel.

? Text Exercise 2.6.5

Consider the spread of the two data sets and compute the MAD from the mean.

$$\{-2, -2, 2, 2\}$$
$$\{-3, -1, 1, 3\}$$

Answer

Our first intuition is to look at the ranges, 4 and 6, and subsequently say that the second data set is more widely dispersed. Both data sets have a mean of 0. The MAD from the mean for the first data set is thus $\frac{2+2+2+2}{4} = 2$, and MAD from the mean for the second data set is $\frac{3+1+1+3}{4} = 2$. So both data sets have the same MAD from the mean. Note that every observation in the first data set is 2 units away from the mean, but in the second data set, two are closer than the average distance, and two are farther than the average distance. The MAD from the mean fails to distinguish between these two sets.

Variance

Recall that the difference between deviation and absolute deviation can be understood as the difference between displacement and distance. An advantage of using distance is that values do not cancel when summing. Using absolute values computes distances in a single dimension. There are many notions of distance in higher dimensions; hopefully, we are familiar with our standard two-dimensional distance formula $d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. Without going into the details, we can understand our next measures of dispersion from the perspective of a higher dimensional distance between our central value v and our data set $x_1, x_2, \ldots, x_n$.

$$d = \sqrt{(x_1 - v)^2 + (x_2 - v)^2 + \cdots + (x_n - v)^2}$$

Note that we sum over all the deviations from v and that each deviation from v is being squared. The details here are beyond the scope of this course, but hopefully, we have built at least an initial intuition as to why we might now consider the **squared deviation from v.** We do so using the same data from the previous section.

Table 2.6.3: An example of various squared deviations

Values	Squared deviations from the mean	Squared deviations from the median
2	$(2-7)^2 = (-5)^2 = 25$	$(2-4)^2 = (-2)^2 = 4$
3	$(3-7)^2 = (-4)^2 = 16$	$(3-4)^2 = (-1)^2 = 1$
4	$(4-7)^2 = (-3)^2 = 9$	$(4-4)^2 = (0)^2 = 0$
9	$(9-7)^2 = (2)^2 = 4$	$(9-4)^2 = (5)^2 = 25$
17	$(17-7)^2 = (10)^2 = 100$	$(17-4)^2 = (13)^2 = 169$
Sum	154	199
Mean	$\frac{154}{5} = 30.8$	$\frac{199}{5} = 39.8$

Interpreting the sums and means of these squared deviations is more complicated. We were dealing with displacements and distances previously; now, we have squared distances. If we had units attached to our data, such as cm, the units on these measures would be units^2,

such as cm^2, but the intuition that we have been building remains consistent. Larger values of these measures indicate greater degrees of spread. We shall encounter an associated measure with a more intuitive interpretation soon.

When considering the MAD, the median minimized the sum of the absolute deviations. We note that with the squared deviations, the median does not minimize the sum because the sum for the squared deviations from the mean is smaller! Indeed, the sum of the squared deviations from the mean is the smallest possible (if you have a background in calculus, see if you can show that the mean minimizes the sum of squared deviations). For this reason, among many others, we define the variance of a population data set as the average of the squared deviations from the mean.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

We denote the variance with σ^2, where σ is the lower case Greek letter sigma. Recall that using Greek letters for a descriptive statistic indicates a population parameter. This is indeed the case here; this formula is for **population variance**.

? Text Exercise 2.6.6

Consider the quiz data from the beginning of this section in the table below. We are only interested in the performance of these particular students and, therefore, treat the data as population data. Compute the variance for both Quiz 1 and Quiz 2.

Table 2.6.4: Scores from Quiz 1 and Quiz 2.

Quiz 1	5	5	6	6	6	6	6	6	7	7	7	7	7	8	8	8	8	9	9	9
Quiz 2	4	4	5	5	5	5	6	6	6	7	7	7	8	8	9	9	9	10	10	10

Answer

To find the population variance, we follow these steps:

1. List each data value.
2. Calculate the mean.
3. Calculate the deviation from the mean for each score.
4. Square the deviations from the mean.
5. Average the squared deviations from the mean.

Table 2.6.5: Calculation of variance for Quiz 1 scores

Quiz 1 **Scores**	**Deviations from the Mean**	**Squared Deviations**
9	$9-7=2$	$2^2=4$
9	2	4
9	2	4
8	$8-7=1$	$1^2=1$
8	1	1
8	1	1
8	1	1
7	$7-7=0$	$0^2=0$
7	0	0
7	0	0
7	0	0
7	0	0
6	$6-7=-1$	$(-1)^2=1$
6	-1	1
6	-1	1

6	-1	1
6	-1	1
6	-1	1
5	$5-7=-2$	$(-2)^2=4$
5	-2	4
$\mu=\frac{\sum x}{N}=7$		$\sigma_1^2=\frac{\sum(x-\mu)^2}{N}=\frac{30}{20}=1.5$

Table 2.6.6: Calculation of variance for Quiz 2 scores

Quiz 2 **Scores**	**Deviations from the Mean**	**Squared Deviations**
10	$10-7=3$	$3^2=9$
10	3	9
10	3	9
9	$9-7=2$	$2^2=4$
9	2	4
9	2	4
8	$8-7=1$	$1^2=1$
8	1	1
7	$7-7=0$	$0^2=0$
7	0	0
7	0	0
6	$6-7=-1$	$(-1)^2=1$
6	-1	1
6	-1	1
5	$5-7=-2$	$(-2)^2=4$
5	-2	4
5	-2	4
5	-2	4
4	$4-7=-3$	$(-3)^2=9$
4	-3	9
$\mu=\frac{\sum x}{N}=7$		$\sigma_2^2=\frac{\sum(x-\mu)^2}{N}=\frac{78}{20}=3.9$

The variance for Quiz 1 is 1.5 and the variance for Quiz 2 is 3.9. From the histograms, we knew that Quiz 2 was more spread out, which we see in Quiz 2 having a larger variance.

? Text Exercise 2.6.7

Recall the two data sets from Text Exercise 2.6.5 (treat them as population data), which were indistinguishable using MAD from the mean. Compute the variance for each data set and explain how the variance can distinguish them while the MAD from the mean cannot.

$$\{-2,-2,2,2\}$$
$$\{-3,-1,1,3\}$$

Answer

Recall that both sets have a mean of 0. So, the squared deviations from the mean are just the values squared. The variance of the first data set is $\sigma_1^2 = \frac{4+4+4+4}{4} = 4$, and the variance of the second data set is $\sigma_1^2 = \frac{9+1+1+9}{4} = 5$. The two data sets are different using variance because the squaring action puts greater weight on values farther from the mean. While -3 and -2 are just one unit away from each other, the -3 contributes 9 to the sum in the variance while the -2 contributes only 4. And similarly, even though we did not see this in our example, if deviations are small, less than 1, their weight is even less by squaring.

We are more interested in the larger population when we have sample data. We use sample statistics to estimate population parameters. Statisticians have found that when calculating the average of squared deviations from the mean using sample data, the computation tends to underestimate the variance of the larger population significantly. Thinking intuitively, this makes sense as there is usually greater variability in a large group than in some subset of that group. Imagine the population was $1,2,3,4,5,6,7,8,9,10$; take a sample from this population, say, 2, $3,5,7$. Notice the sample is less dispersed than the population. The sample will, more often than not, have less variance than the population. If we want to estimate the population variance based on a sample, using a number larger than the value obtained from the above formula is better. Because of this, statisticians have adjusted the averaging process when dealing with sample data to result in a better inferential measure; the solution is to divide by $n-1$ rather than n. The **sample variance** is defined as follows.

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Notice the use of s, a Latin letter, in s^2; the Latin letter again reminds us that we are dealing with sample data. In practice, variance is usually computed in a sample, so this formula is often used.

? Text Exercise 2.6.8

Let's take a concrete example. Consider a random sample of 10 quiz scores from Quiz 2. We constructed a random sample from the table on the previous problem: $\{5,5,5,6,7,7,7,8,10,10\}$. Calculate the sample variance for this sample data. Compare the sample variance to the population variance.

Answer

$$\bar{x} = \frac{5+5+5+6+7+7+7+8+10+10}{10} = \frac{70}{10} = 7$$

$$s^2 = \frac{[(5-7)^2+(5-7))^2+(5-7))^2+(6-7))^2+(7-7))^2+(7-7))^2+(7-7))^2+(8-7))^2+(10-7))^2+(10-7))^2]}{(10-1)}$$

$$= \frac{(4+4+4+1+0+0+0+1+9+9)}{9}$$

$$= \frac{32}{9} = 3+\frac{5}{9}$$

$$\approx 3.5556$$

The population variance σ^2 is 3.9 while the sample variance s^2 is about 3.5556.The values are off by about 0.3444,which is much closer than it would have been if we had divided by 10 rather than 9 in our computation. Without adjusting the average for sample variance, we would compute 3.2,which is 0.7 away from the population variance.

Standard Deviation

Variance is a powerful measure and is the basis for much of statistics. We struggle to interpret this value because of the squared units. Consequently, we introduce another measure closely related to the variance: the **standard deviation**. The standard deviation is simply the square root of the variance. The population standard deviation is the square root of the population variance, and the sample standard deviation is the square root of the sample variance. This makes the units on the measure the same as those of the original data; for example, if the original data was in cm, then the standard deviation will also be measured in cm. We can understand the standard deviation loosely as a measure of the distance a typical value is from the mean. Our natural choices of symbols are the bases for our two variances: σ for population and s for sample.

? Text Exercise 2.6.9

Compute the standard deviations for both Quiz 1 and Quiz 2.

Answer

Since we have already computed the variances for Quiz 1 and Quiz 2 in text exercise 2.6.5.We only need to take the square root of each.

$\sigma_1 = \sqrt{1.5} \approx 1.225$

$\sigma_2 = \sqrt{3.9} \approx 1.975$

? Text Exercise 2.6.10

What is the smallest value that standard deviation can take on? Construct a data set of 5 observations with such a standard deviation.

Answer

As the square root of variance, we must analyze the square root function and variance as a measure. Square roots return nonnegative values. The smallest nonnegative value is 0. The square root is equal to 0, $\sqrt{d} = 0$, only when the input d is 0. Can the variance of a data set be 0? An average is 0 only when the sum of the values is 0. We are adding up squared deviations, which are all nonnegative. The only way a sum of nonnegative values is 0 is if they are all 0. The deviation from the mean is 0 only when the observation is equal to the mean. There is no variability in the data values. An example of such a data set would be $\{1,1,1,1,1\}$.

It is worth noting that all of the measures discussed here (range, IQR, MAD, variance, standard deviation) are always non-negative. Moreover, they will all be 0 on constant data sets, like $\{1,1,1,1,1\}$; conversely, the range, MAD, variance, and standard deviation will not be 0 if there are at least two different data values.

2.7: Distributions- Using Centrality and Variability Together

Learning Objectives

- State and apply Chebyshev's Inequality
- Define unusual observations
- Define normal distributions
- State properties of normal distributions
- Discuss distributions and curves
- Define the standard normal distribution
- State and apply the Empirical Rule
- Define z-score
- Define outliers

Section 2.7 Excel File (contains all of the data sets for this section)

Connecting Measures of Central Tendency & Measures of Dispersion

In the previous two sections, we developed two significant classes of descriptive statistics: measures of central tendency and measures of dispersion. In this section, we begin to consider the power of these measures together. We have shown that the mean of a data set is the balancing point of its frequency distribution and that it minimizes the sum of the square deviations. In the last section, we defined the variance of a data set to be the mean of these square deviations (with appropriate modifications for sample data). We set the standard deviation to be the square root of the variance. The coupling of these measures of centrality and dispersion tells us a lot about the distribution of our data.

Think about what "standard deviation" means; it represents a measure of "typical distance" from the mean. If the mean of some data set was 100 and the standard deviation was 10, then we would expect a good chunk of our data points to be between 90 and 110; that is, much of the data does not deviate from the mean by more than the standard deviation. We would expect most of the data to fall between 80 and 120; that is, most of the data would be within two standard deviations of the mean. Think about it: if all of our data points were less than 80 or more than 120, then all of them deviate from 100 by at least 20. How could a "typical" deviation from the mean be 10 if most of the points are off by at least 20? Taking this further, we should find it incredibly rare that a data point is more than 7 standard deviations away from the mean.

? Text Exercise 2.7.1

Consider the data set $\{2,3,4,14,15,16,16,17,27,36\}$.

1. Give the population mean and population standard deviation.

Answer

Using the formulae from previous sections, the mean is $\mu = 15$ and the standard deviation is $\sigma \approx 10.13$.

2. How many data points are within 1 standard deviation of the mean?

Answer

For a data point to be within 1 standard deviation of 15 means that its distance to 15 is no more than 10.13.Thus, any data point between $15 - 10.13 = 4.87$and $15 + 10.13 = 25.13$would be within 1 standard deviation of the mean. We can see that 5 data points fall in this range, meaning $50\%$of the data points are within 1 standard deviation of the mean.

3. What proportion of data points are within 1.5 standard deviations of the mean?

Answer

1.5 standard deviations would be $1.5\sigma \approx 1.5 \cdot 10.13 \approx 15.2$. Thus, any number whose distance to the mean is less than $15.2$is within 1.5 standard deviations of the mean. If we go 15.2below the mean, that would be $15 - 15.2 = -0.2$. If we go 15.2above the mean, that would be $15 + 15.2 = 30.2$. Notice that all but 1 of our data points fall in this range. Since we have 10 data points, that yields $90\%$of our data is within 1.5 standard deviations of the mean.

4. What proportion of data points are within 3 standard deviations of the mean?

Answer

$3\sigma \approx 3 \cdot 10.13 = 30.39$. Notice that none of the data points differ from the mean by more than 30.39. This means all of our data is within 3 standard deviations of the mean. The proportion is 100%.

We note that the behavior in the preceding example does not characterize all data sets. Generally, we can have data sets where some points are 5, 10, or any number of standard deviations above or below the mean. These examples have a lot of data points; however, the basic intuition still stands: only a tiny proportion of points are far away from the mean. Let us be more precise with this statement.

Chebyshev's Inequality

Around the middle of the nineteenth century, mathematicians (Pafutny Chebyshev, in particular) discovered an explicit connection between a data set's mean and standard deviation and its distribution. The explicit development of such a result is beyond the scope of an elementary statistics course, so we shall present the result and begin to digest the implications.

Chebyshev's Inequality

Given any data set with (population) mean, μ, and (population) standard deviation, σ, and any real number $k > 1$, the proportion of observations that lie in the interval $[\mu - k\sigma, \mu + k\sigma]$ is at least $1 - \frac{1}{k^2}$.

Using Chebyshev's Inequality, we can guarantee a minimum percentage of observations falling in an interval symmetric about the mean. By starting at the mean and going a specified number of standard deviations above the mean and then below the mean, we are guaranteed to catch at least a certain percentage of the observations in the data set. This result's great power and beauty come from this inequality being valid for all data sets. That is important to remember. Consider the following basic applications of the result:

If $k = 2$, we have that at least $1 - \frac{1}{2^2} = \frac{3}{4} = 75\%$ of the observations fall between $\mu - 2\sigma$ and $\mu + 2\sigma$. Another way to say this is that, for any data set, at least 75% of the data falls within two standard deviations of the mean.

If $k = 3$, we have that at least $\frac{8}{9} = 88.\bar{8}\%$ of the observations fall between $\mu - 3\sigma$ and $\mu + 3\sigma$.

The implication of this is that for any data set, less than 25% of the observations fall more than 2σ away from the mean μ and less than $11.\bar{1}\%$ of the observations fall more than 3σ away from the mean μ. Most observations fall within 2 or 3 standard deviations from the mean. When we have an observational value that falls away from the bulk of the observations, we consider it unusual. We say that an observation is unusual by the 2 standard deviation rule if it is more than 2 standard deviations away from the mean; likewise, an observation is unusual by the 3 standard deviation rule if it is more than 3 standard deviations away from the mean.

? Text Exercise 2.7.2

Suppose a sales department of some corporation is supposed to acquire a minimum of $\$200,000$ in revenue each week. Glancing at a long-term report over the last 25 years, we see that, on average, the department made $\$245,000$ each week with a population standard deviation of $\$20,000$. What can be said about how often the department did not meet the quota?

Answer

Notice that the quota is $\$45,000$ below the average revenue generated. We need to know how many standard deviations equal $\$45,000$ to apply Chebyshev's Inequality. Since the standard deviation is $\$20,000$, we can divide to obtain this.

$$\frac{45000}{20000} = \frac{45}{20} = 2 + \frac{1}{4} = 2.25$$

Each week the department did not meet the quota was at least 2.25 standard deviations below the mean. Chebyshev's Inequality guarantees, on any data set, that the proportion of data points within 2.25 standard deviations of the mean is at least $1 - 1/(2.25)^2 \approx 0.80$. We can be confident that at least 80% of the time, the department met the quota. It is possible that they met the quota far more than 80% (they could have met it 100% of the time). We would need more information to

obtain a more precise estimate. Regardless, we can be sure that the department did not miss quota more than 20% of all weeks in the last 25 years.

? Text Exercise 2.7.3

Using Chebyshev's Inequality, determine the number of standard deviations from the mean k to guarantee at least 50% of the observations to be in the interval $[\mu - k\sigma, \mu + k\sigma]$.

Answer

We first note that $50\% = \frac{1}{2}$. Our problem reduces to solving the following equation:

$$\frac{1}{2} = 1 - \frac{1}{k^2}$$

Meaning that

$$\frac{1}{2} = \frac{1}{k^2}$$

Which yields the solution:

$$k = \pm\sqrt{2}$$

Remember, k can be any real number greater than 1. Hence $k = \sqrt{2}$.

? Text Exercise 2.7.4

Can we use the results of the previous exercise to compute the first and third quartiles for any data set? Explain.

Answer

There are at least two reasons why we cannot do this. While 50% of the observations do fall between the first and third quartiles, it is also true that 25% fall below the first quartile and 25% fall above the third quartile. Chebyshev's Inequality does not guarantee that the percentage of observations in each tail is 25%.

Chebyshev's Inequality asserts a minimum percentage of observations in the interval. It does not claim that there are exactly 50% of the observations; it argues that there are at least 50%.

It is worth noting that Chebyshev's Inequality does tell us that the first quartile cannot be more than 2 standard deviations below the mean, as this would imply that more than 75% of the data is larger than the first quartile. Similarly, the third quartile cannot be more than 2 standard deviations above the mean.

Normal Distributions and Curve Fitting

While Chebyshev's Inequality is powerful because it applies to all distributions, more precise connections can be made when we restrict our interest to particular classes of distributions. We shall encounter several classes throughout our course of study, but at this point, we shall limit ourselves to **normal distributions**. Normal distributions are common in data from everyday life; heights, IQ scores, and the "bell curve" of class grades are familiar examples. Normal distributions are symmetric and unimodal, with the mean, median, and mode all equal.

To uniquely express the shape of a normal distribution, we must discuss modeling distributions with mathematical functions or curves. We can use continuous functions to model both discrete and continuous data. Consider the following figure.

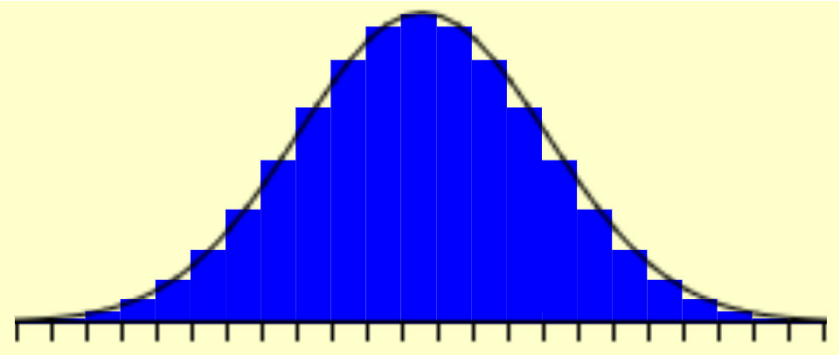


Figure 2.7.1: Continuous function fit to a histogram

The curve highlights the shape of the histogram reasonably well and could continue to fit better if the histogram had more classes. Increasing the number of classes is not always possible with discrete variables and finite data sets. Still, it could happen with continuous variables provided enough data is available with sufficient measurement precision. Recall that frequency and relative frequency distributions have similar graphical representations; the only differences are in the vertical scales. As such, we could develop functions using either distribution. We consider relative frequency distributions; these curves will play an important role throughout this course.

At each point along the horizontal axis, we have two values to compare vertically: the height of the bar versus the function value. The height of the bar represents the percentage of observations that fall in that class. We want to be able to retain this information with our model (function). As we can see, the value of the function changes within classes, making retaining this information difficult. Our solution is to construct curves that closely resemble common classes of histograms so that the area underneath the curve over a given interval corresponds to the relative frequency of the class(es) in that interval.

Consider this process visually using a data set from Statistics Online Computation Resource containing $25,000$ height values accurate to 5 decimal places. We construct relative frequency distributions with class widths of 1, 0.1, and 0.01 and portray them in two ways graphically. The histograms on the left represent the relative frequency of a class using the height of its bar. In contrast, the graphical representations on the right represent the relative frequency of a class using the area of its bar. Note that the vertical scales remain the same across all 6 graphs.

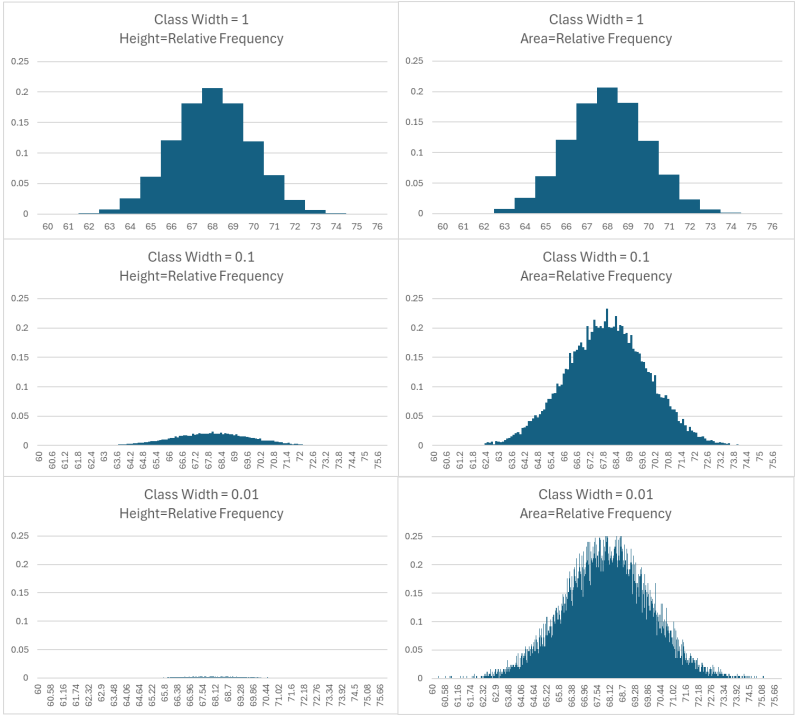


Figure 2.7.2: Graphical representations of relative frequency by height (left) and area (right)

Since we have a finite data set, the relative frequencies of each class become extremely small, around $\frac{1}{25000}$, as the class widths become smaller. We see each class's height get smaller until it is difficult to see (bottom left graph). We see a different story on the right. Since our relative frequencies are represented by the area of the bars and the class widths are getting smaller, the shape of the distribution seems to solidify as our class widths decrease. In taking smaller and smaller class widths, our graphical representation becomes "smoother" in the shape of a continuous function, and the area underneath the function over an interval corresponds to the relative frequency of the observations in that interval.

A significant component of statistical research is checking how closely any particular model fits our actual data set. Continuous models allow us to build our statistical framework around these functions using the power of mathematics without needing to construct something new for every data set we study.

Our chosen models preserve relative frequency through area. The relative frequency of observations over a given interval is the area under the curve over that same interval. Recall that the sum of all the relative frequencies of a distribution is always equal to 1; this means that the area underneath the entirety of these curves will also be 1. We will name these curves and continue to deepen our understanding in the coming chapters.

With all of this build-up, we are now ready to define the class of normal distributions; the curve that defines them depends on two factors: the mean and standard deviation. While the knowledge of the particular function bears little utility in this course, we now provide it with the general normal distribution graphed below.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

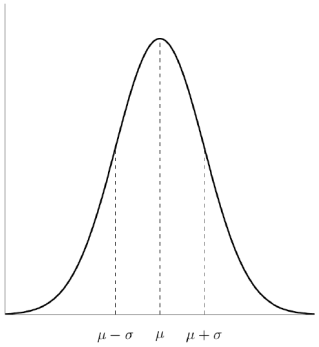
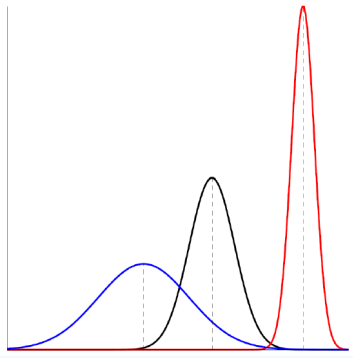


Figure 2.7.3: The normal distribution centered at μ with standard deviation σ

Text Exercise 2.7.5

Use the figure to answer the following:

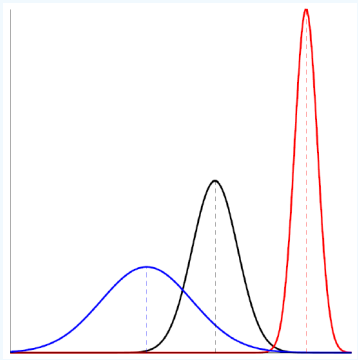
1. The means of these normal distributions are -3, 0, and 4. Determine to which distribution each value belongs.



Answer

Given that the mean, median, and mode are all equal. The mean occurs at the peak of the distribution. Thus $\mu_{\text{blue}} = -3$, $\mu_{\text{black}} = 0$, and $\mu_{\text{red}} = 4$.

2. The standard deviations of these normal distributions are $0.5, 1$, and 2. Determine to which distribution each value belongs.



Answer

The standard deviation is a measure of dispersion. The smaller the spread, the smaller the standard deviation. Since the blue distribution is spread out the most, the blue distribution has the largest standard deviation. We can say $\sigma_{\text{blue}} = 2$, $\sigma_{\text{black}} = 1$, and $\sigma_{\text{red}} = 0.5$.

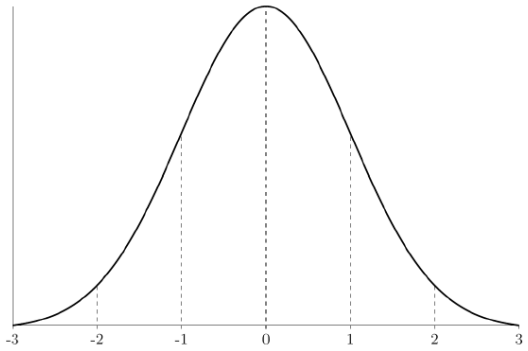


Figure 2.7.4: Standard Normal Distribution $\mu = 0$ and $\sigma = 1$

The normal distribution shown above is called the **standard normal distribution** or the **z-distribution.** The standard normal distribution is the normal distribution with $\mu = 0$ and $\sigma = 1$. We can quickly tell that the mean of the distribution is 0. There is also a way to determine the standard deviation; the reasoning behind it is less apparent, but a first-semester calculus student should be able to arrive at the conclusion. There are two inflection points on normal curves, and they happen at exactly one standard deviation below and one standard deviation above the mean. All we need to do is identify an inflection point and determine the distance to the mean. For those who do not know what inflection points are, look around the points -1 and 1 in the figure above to see what is happening; note that these values are one standard deviation away from the mean. We notice that between -1 and 1, the curve seems to open downward, and along the tails, the curve appears to open upward. At some point, the function switches from opening upward to downward, and then at another point, the function switches from opening downward to upward; these points are called inflection points. In a normal distribution, the inflection points always occur at one standard deviation above and below the mean.

The Empirical Rule

We began our discussion about normal distributions by saying claims stronger than Chebyshev's Inequality can be made when we restrict our distributions to particular classes. We now formulate such a result for normal distributions, which we call the Empirical Rule.

The Empirical Rule

Given a normal distribution with mean μ and standard deviation σ, the percentage of observations within 1, 2, and 3 standard deviations of the mean is known

$$\% \text{ in } [\mu - \sigma, \mu + \sigma] \approx 68\%$$
$$\% \text{ in } [\mu - 2\sigma, \mu + 2\sigma] \approx 95\%$$
$$\% \text{ in } [\mu - 3\sigma, \mu + 3\sigma] \approx 99.7\%$$

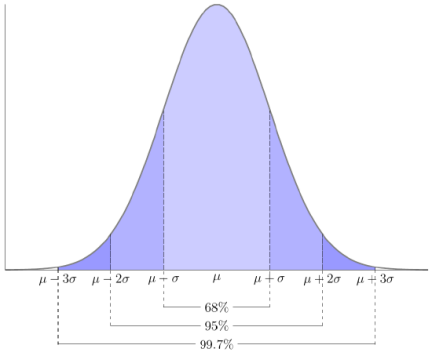


Figure 2.7.4: The Empirical Rule

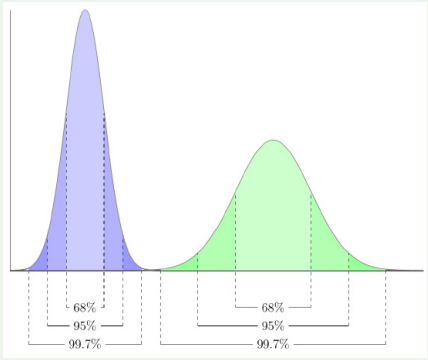


Figure 2.7.5: Empirical Rule with two normal distributions with different means and standard deviations on the same set of axes

Note: the approximation signs in the statement of the Empirical Rule are used because the areas underneath the curves over the appropriate intervals can be approximated to many decimal places. However, we do not expect that sort of precision at this stage. In future chapters, we will use technology for greater accuracy.

Notice the difference between the claims of Chebyshev's Inequality and the Empirical Rule. Chebyshev's Inequality provides a lower bound for the percentage of observations within k standard deviations of the mean for any data; meanwhile, the Empirical Rule asserts what those percentages are for 1, 2, and 3 standard deviations, but only for data that is normally distributed.

? Text Exercise 2.7.6

Let us revisit a previous example. Suppose a sales department of some corporation is supposed to acquire a minimum of $\$200,000$ in revenue each week. Glancing at a long-term report over the last 25 years, we see that, on average, the department made $\$245,000$ each week with a population standard deviation of $\$20,000$. Suppose that we also know that the data is normally distributed. What can be said about how often the department did not meet the quota?

Answer

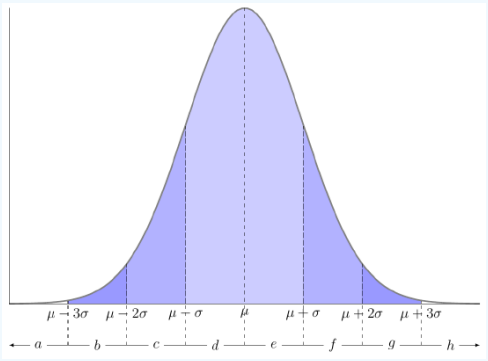
Last time, we noticed that the quota was 2.25 standard deviations below the mean and that Chebyshev's Inequality guaranteed at least 80% of the data points were above $\$200,000$. Now we have more information: we are given that the data is normally distributed; therefore, we can be more precise. The Empirical Rule tells us that 95% of the data is no more than 2 standard deviations away from the mean. Two standard deviations is $\$40,000$, so we are saying 95% of the data falls between $245,000 - 40,000 = 205,000$ and $245,000 + 40,000 = 285,000$. Therefore, at least 95% of our data is above quota. We can use the symmetry of the normal distribution to say even more.

The Empirical Rule tells us only 5% of the data is more than 2 standard deviations away from the mean. Because the curve is symmetric, half lies above the mean and half below the mean. Anything above the mean was above the quota; the 2.5% of the data points more than 2 standard deviations above the mean can be added to the values above the quota. We conclude that at least $97.5\%$of the data was above the quota. The department missed the quota no more than $2.5\%$of the time. Later, we will develop tools that will allow us to be even more precise.

We can also return to our ideas regarding unusual observations. When we are beyond 2 or 3 standard deviations away from the mean for normal distributions, the percentage of observations that lie there are 5% or 0.3% respectively. Here the title of unusual**,** rings a little stronger.

? Text Exercise 2.7.7

Use symmetry and the Empirical Rule to find the percentage of observations in each of the following intervals for a normal distribution with mean μ and standard deviation σ.



1. $(-\infty, \mu - 3\sigma]$

Answer

$(-\infty, \mu - 3\sigma] \approx 0.15\%$

2. $[\mu - 3\sigma, \mu - 2\sigma]$

Answer

$[\mu - 3\sigma, \mu - 2\sigma] \approx 2.35\%$

3. $[\mu - 2\sigma, \mu - \sigma]$

Answer

$[\mu - 2\sigma, \mu - \sigma] \approx 13.5\%$

4. $[\mu - \sigma, \mu]$

Answer

$[\mu - \sigma, \mu] \approx 34\%$

5. $[\mu, \mu + \sigma]$

Answer

$[\mu, \mu + \sigma] \approx 34\%$

6. $[\mu + \sigma, \mu + 2\sigma]$

Answer

$[\mu+\sigma, \mu+2\sigma] \approx 13.5\%$

7. $[\mu+2\sigma, \mu+3\sigma]$

Answer

$[\mu+2\sigma, \mu+3\sigma] \approx 2.35\%$

8. $[\mu+3\sigma, \infty)$

Answer

$[\mu+3\sigma, \infty) \approx 0.15\%$

Text Exercise 2.7.8

IQ scores are generally thought to be normally distributed with a mean of 100 and standard deviation of 16. Determine the percentage of the population with IQ scores in the given ranges.

1. Between 84 and 116

Answer

Since 84 is 16 less than 100, 84 corresponds to 1 standard deviation below the mean. Likewise, 116 is 1 standard deviation above the mean. A direct application of the Empirical Rule tells us that 68% of the population is within this range.

2. Between 84 and 132

Answer

Since 132 is 32 greater than 100 and $\frac{32}{16} = 2$, 132 lies 2 standard deviations from the mean. We are looking at the interval from 1 standard deviation below the mean to 2 standard deviations above the mean. The previous exercise shows that this range contains $68\% + 13.5\% = 81.5\%$ of the population.

3. Greater than 148

Answer

Since 148 is 48 greater than 100 and $\frac{48}{16} = 3$, 148 lies 3 standard deviations from the mean. The percentage of the population that lies beyond that is 0.15%.

4. Between 52 and 68

Answer

Since 52 is 48 less than 100, 52 is 3 standard deviations below the mean. Likewise, 68 is 2 standard deviations below the mean. The percentage of the population that lies between 52 and 68 is 2.35%.

5. Explain why we cannot determine the percentage of the population between 100 and 108 using the Empirical Rule and symmetry.

Answer

It might be tempting to say that the percentage of the population between 100 and 108 is 17% because we have often split the percentages evenly across our known intervals. We cannot do this because we do not have symmetry over the interval $[\mu, \mu+\sigma]$. The area under the curve from 100 to 108 is larger than the area under the curve from 108 to 116. In future chapters, we will use technology to compute the area and thus deduce the percentage of the population within such intervals.

z-scores

Notice that in both the Empirical Rule and Chebyshev's Inequality, we are interested in how many standard deviations an observation is from the mean. In the previous exercise, we repeatedly determined how far away an observation was from the mean. Then, we divided that difference by the standard deviation to determine the number of standard deviations the observation was from the mean. This computation is commonly called a "**standardization of the data**" and is known as an observation's **z-score**.

$$z = \frac{x - \mu}{\sigma}$$

Our z-scores do more than facilitate Empirical Rule calculations; as "standardized" measures, they enable us to compare observational values across different populations.

Text Exercise 2.7.9

As a married couple prepared to send their daughter to college in 2017, they wanted to compare relative high school academic prowess. The daughter only took the SAT. The mom and dad took the ACT, but there was an age gap of several years. The dad took the ACT in 1995 while the mom took the ACT in 1999. After doing a little research, they found out that the average score on the ACT in 1999 was 21 with a standard deviation of 4.7 and in 1995 the average score was 20.8 with a standard deviation of 4.7. The average score on the SAT in 2017 was 1060 with a standard deviation of 195. Determine who achieved the highest relative academic prowess on standardized tests if the dad earned a 29, the mom earned a 28, and the daughter earned a 1395 on their respective exams.

Answer

In comparing their values, we can see the dad barely outscored the mom. However, we cannot directly compare the daughter's score as the scale for the SAT is entirely different from the scale for the ACT. One way to compare the observed values in these three separate populations is to compute and then compare each observation's z-score.

$$z_{\text{dad}} = \frac{29 - 20.8}{4.7} \approx 1.745$$
$$z_{\text{mom}} = \frac{28 - 21}{4.7} \approx 1.489$$
$$z_{\text{daughter}} = \frac{1395 - 1060}{195} \approx 1.718$$

Based on the z-scores, the dad performed the best, followed closely by his daughter. We might also consider whether these values are significantly different from each other. That is, the dad's z-score was (\ 0.027 \) larger than the daughter's z-score...is such a difference meaningful? We will answer these questions in future work once more measurements are developed.

Unusual Observations and Outliers

As we have progressed through this section, we have referenced the idea of unusual observations twice and mentioned that there is no standard definition agreed upon by all professionals. Chebyshev's Inequality allows us to estimate the minimum percentage of observations within a certain number of standard deviations of the mean. The Empirical Rule only makes assertions about the percentage of observations in normal distributions. From these two results, we know we have a very small percentage of observations, many standard deviations away from the mean; we classify such observations as unusual**.** Sometimes, we have a few isolated observations positioned far from the rest of our data, called **outliers**. Outliers can point to rare/unique occurrences or possibly measurement errors. When an outlier is present, we want to check the validity of the measurement. If protocols were violated or an error occurred in the measurement, we will likely remove the observation from our data analysis.

Text Exercise 2.7.10

If an observation is considered unusual by the 2 standard deviation rule, what can we say about its z-score?

Answer

Since the observation is considered unusual by the 2 standard deviation rule, we know it lies at least 2 standard deviations away from the mean. The z-score is the number of standard deviations an observation is from the mean. We know the magnitude of the z-score is at least 2. It could be negative or positive.

One way to classify outliers is using box plots. The box contains 50% of the observations. How far must an observation be outside this box to be classified as an outlier? Recall the interquartile range IQR, a measure of dispersion, a range measure of the middle 50% of our ordered data. The box represents our central data region, and the IQR is the length of the box. It is common practice to say any observation beyond the box by more than $1.5 \cdot \text{IQR}$ is an outlier.

? Text Exercise 2.7.11

Using the 30 scores from the 10 point assignment in section 2.1, determine if there are any unusual observations or outliers. Use both of the rules for determining unusual observations.

$$\{3, 4, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10\}$$

Answer

The rules about unusual observations depend on the mean and the standard deviation. We are studying this data as population data. A quick computation gives us the following values $\mu = 7\frac{2}{15} \approx 7.267$ and $\sigma \approx 1.769$. Since these are intermediary steps to our conclusion, we do not want to use the rounded values in future computations. Reference them exactly when using technology.

The first unusual observation rule is if the observation is beyond 2 standard deviations from the mean. The bounds for this are lower bound $= \mu - 2 \cdot \sigma \approx 7.267 - 2 \cdot 1.769 \approx 3.729$ and upper bound $= \mu + 2 \cdot \sigma \approx 7.267 + 2 \cdot 1.769 \approx 10.804$. By this standard, the single data value 3 is considered unusual.

The second unusual observation rule is if the observation is beyond 3 standard deviations from the mean. The bounds for this standard are lower bound ≈ 1.96 and upper bound ≈ 12.573. By this standard, there are no unusual observations.

The outlier rule depends on Q_1 and Q_3. $Q_1 = 6$ and $Q_3 = 9$. The $\text{IQR} = 9 - 6 = 3$ and $1.5 \cdot \text{IQR} = 4.5$. The bounds for this standard are lower bound $= Q_1 - 1.5 \cdot \text{IQR} = 6 - 4.5 = 1.5$ and upper bound $= Q_3 + 1.5 \cdot \text{IQR} = 9 + 4.5 = 13.5$. Since no observations fall outside of this interval, there are no outliers.

2.8: Measures of Median and Mean on Grouped Data

Learning Objectives

- Determine the median and mean in grouped discrete data
- Determine the median and mean in grouped relative frequency data
- Determine the median and mean in grouped continuous data
- Extend to weighted mean

Section 2.8 Excel File (contains all of the data sets for this section)

Introduction to Grouped Data

In our investigation of descriptive statistics, we worked on a collection of individual data values and then formed appropriate summary measures of that "raw" data. However, we may sometimes be given the data in a summarized frequency distribution format instead of as a "raw" data collection. Can we find our various descriptive statistic measures if we only have the frequency table, which represents the data in grouped form? Although the answer is only "sometimes," the underlying concept of how we can do so is essential for later ideas in the course. We begin with mean and median measures found from frequency tables on grouped quantitative data, but where the grouping was not formed by interval values but only by the same single values.

Mean and Median of Non-Interval but Grouped Data in a Frequency Table

Look at the frequency distribution in Table $2.8.1$ shown below from Section 2.1 about thirty student scores (discrete 10-point scale) for an assignment; we will assume these are only a sample of a larger population of scores. Notice that our table shows no loss of crucial information on the data as each distinct data value is explicitly shown in the table, and no intervals are used to represent the grouped student scores.

Table 2.8.1: Grouped Frequency Distribution

Student Score	Frequency
3	1
4	1
5	3
6	5
7	5
8	7
9	5
10	3

With such a table, we could formally recreate the entire data set $\{3,4,5,5,5,\ldots,9,10,10,10\}$ by recognizing the meaning of the frequency values for each of the various score values in the table. If we had more data values, recreating the data set would be tedious, and we could lose information on the data. Even if using technology to produce our descriptive measures, we must "type in" all individual data values. This process is likely to lead to many data entry errors. Recreating the "raw" data set is unnecessary; we can determine the median and the arithmetic mean by working with data in this frequency distribution format.

First, we examine the median measure by using quantitative reasoning. We note by the sum of the frequency column that there are 30 pieces of data and, by our earlier discussion in Section 2.4, the median is the average of the two data values in position 15 and 16 in the ordered list of all data values. Using our frequency column, we accumulate across our frequency counts to see that the 15^{th} data position is within the group of "7" scores and the 16^{th} data position is within the group of "8" scores (total accumulation of number of data scores from "3" to "7" includes $1+1+3+5+5 \ = 15$ scores). So the median value of the data is $\frac{7+8}{2} = 7.5$. In general, we can find the median by focusing on our frequency counts to help us determine the center position location, using the location value to determine the median within the grouped variable values.

Next, we examine the mean. Recall that the mean is found in our original discussion by summing our quantitative data values, then dividing by the number of values in the data set: $\bar{x} = \frac{\sum x_i}{n}$. Notice in our grouped data, we can find the portion of the entire sum generated by each group by multiplying the group data value by the frequency. For example, the grouped data value 5 will contribute a total of $3 \cdot 5 = 15$ to the total sum since there are three 5 values in the data set; similarly, the grouped data value 8 will contribute a total of $8 \cdot 7 = 56$ toward to total sum since there are seven 8 values in the data set. This leads us to the following adjustment of our table to compute the mean of the grouped data (also note the change to general headings on each column).

Table 2.8.2: Computation of arithmetic mean for data from Table 2.8.1

x_j	f_j	$x_j \cdot f_j$
3	1	3
4	1	4
5	3	15
6	5	30
7	5	35
8	7	56
9	5	45
10	3	30
Totals:	$\sum f_j = 30$	$\sum (x_j \cdot f_j) = 218$
	Arithmetic Mean: $\bar{x} = \frac{\sum (x_j \cdot f_j)}{\sum f_j} = \frac{218}{30} \approx 7.2667$	

In conclusion, by summing our f_j frequency column values, we know the sample size n. We have accomplished the exact computation by adding our thirty individual data values together by summing our $x_j \cdot f_j$ column of values. The arithmetic mean of the data is found by our last computation $\bar{x} = \frac{\sum (x_j \cdot f_j)}{\sum f_j}$. We divided the total sum of all data values by the number of data values. In grouped data of this form, we can find the mean by the above process, and described symbolically by the given formula:

Sample Mean from a Frequency Distribution

$$\bar{x} = \frac{\sum x_i}{n} = \frac{\sum (x_j \cdot f_j)}{\sum f_j}$$

If the data in our table had been population data, we would still perform the same calculation using the same reasoning. And have:

Population Mean from a Frequency Distribution

$$\mu = \frac{\sum x_i}{N} = \frac{\sum (x_j \cdot f_j)}{\sum f_j}$$

? Text Exercise 2.8.1

Consider the Quiz 1 data from section 2.6 below in the frequency table format. Determine the mean from the grouped format and compare it with the results obtained in section 2.6 from the "raw" data. We assume the data is population data in this example.

Table 2.8.3: Grouped Frequency Distribution of Quiz 1 Data

Quiz Scores	Frequency
5	2
6	6

Quiz Scores	Frequency
7	5
8	4
9	3

Answer

To find the mean from this frequency table, we follow these steps with thoughts of what we are doing with the data:

1. Sum the frequency column f_j to determine the size of the data set.
2. Compute the column of values $x_j \cdot f_j$ to weight each quiz score with their occurrence frequency.
3. Sum the column of $x_j \cdot f_j$ values to produce the total sum as if summing the individual data values.
4. Produce the mean by dividing the sum of the $x_j \cdot f_j$ column by the sum of the frequency column f_j.

Table 2.8.4 Computation of arithmetic mean

x_j	f_j	$x_j \cdot f_j$
5	2	10
6	6	36
7	5	35
8	4	32
9	3	27
	$\mu = \frac{\sum x_j \cdot f_j}{\sum f_j} = \frac{140}{20} = 7$	

We notice this is the exact arithmetic mean value computed when working with the twenty individual quiz scores.

Mean and Median of Grouped Data in a Relative Frequency Table

What would happen if we had a relative frequency distribution of this data instead of a frequency distribution table? Recall that relative frequency in this situation measures the proportion of the data set that has a specific data value. We will be using $P(x_j)$ to represent the relative frequency or proportion measure as tied to specific data value x_j.

Table 2.8.5: Relative frequency table of student scores from Table 2.8.1

Student Score x_j	Relative Frequency $P(x_j)$
3	$\frac{1}{30} \approx 0.0333 = 3.33\%$
4	$\frac{1}{30} \approx 0.0333 = 3.33\%$
5	$\frac{3}{30} = 0.1000 = 10.00\%$
6	$0.1667 = 16.67\%$
7	$0.1667 = 16.67\%$
8	$0.2333 = 23.33\%$
9	$0.1667 = 16.67\%$
10	$0.1000 = 10.00\%$
Totals:	$\sum P(x_j) = 1.0000 = 100\%$

With a relative frequency table, we could not formally recreate the entire data set unless we first knew the number of data values in the data set (i.e., the sample or population size). However, we do not need such information to determine the distribution's median or mean. We proceed as above, working with relative frequency measures instead of counted frequency measures.

We first examine the median measure. All data is accounted for by the sum of the frequency column that 100%; we should always sum our relative frequency measures to see if we have a total of $1.0000 = 100\%$. As discussed previously, the median is at the 50^{th} percentile position in the ordered list of our data set. Using our relative frequency column, we can accumulate our relative percentages to see that the 50% data position is right on the border between the group of "7" scores and the group of "8" scores; total relative frequency accumulation of number of data scores from "3" to "7" includes $3.33\% + 3.33\% + 10.00\% + 16.67\% + 16.67\% = 50\%$. The median value of the data is $\frac{7+8}{2} = 7.5$, just as above. We can find the median by focusing on our relative frequency measures to help us determine the location of 50% of the data set from the smallest value and the 50% location value to determine the median within the grouped variable values.

Next, we examine the mean measure. In forming the relative frequency measures, we divided each frequency count by the sample size to form the relative frequency measures. This earlier division and some algebraic reasoning show how we can adjust our standard arithmetic mean formula to fit this situation.

$$\begin{aligned} \bar{x} &= \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} \\ &= \frac{x_1}{n} + \frac{x_2}{n} + \ldots + \frac{x_n}{n} = \sum \frac{x_j \cdot f_j}{n} \\ &= \sum \left(x_j \cdot \frac{f_j}{n} \right) = \sum \left[x_j \cdot P(x_j) \right] \end{aligned}$$

$P(x_j)$ stands for the relative frequency of data value x_j; it is the proportion of the data set with that specific data value, x_j. In a sense, each distinct data value is being "weighted" by the relative frequency of occurrence. For example, the fact that the data value 8 occurs with 23.33% relative frequency should make this data value "weigh-in" more heavily to the average than does the data value 5 that only occurs with 10% relative frequency. Our relative frequency gives us this "weighting" of the data in a relative sense instead of the above, in which the actual frequency measures give us a weighted "count" sense. This leads us to the following adjustment of our table to compute the mean of the grouped data (note the change to general headings on each column).

Table 2.8.6: Computation of arithmetic from relative frequency found in Table 2.8.5

x_j	$P(x_j)$	$x_j \cdot P(x_j)$
3	$\frac{1}{30} \approx 0.0333 = 3.33\%$	$3 \cdot \frac{1}{30} = 0.1000$
4	$\frac{1}{30} \approx 0.0333 = 3.33\%$	$4 \cdot \frac{1}{30} \approx 0.1333$
5	$\frac{3}{30} = 0.1000 = 10.00\%$	$5 \cdot \frac{3}{30} = 0.5000$
6	$0.1667 = 16.67\%$	1.0000
7	$0.1667 = 16.67\%$	1.1667
8	$0.2333 = 23.33\%$	1.8667
9	$0.1667 = 16.67\%$	1.5000
10	$0.1000 = 10.00\%$	1.0000
Totals:	$\sum P(x_j) = 1.0000 = 100\%$	$\sum \left[x_j \cdot P(x_j) \right] \approx 7.2667$

We notice that the results from this relative frequency distribution are the same as those from the previous section of the plain frequency distribution.

In conclusion, by multiplying each unique data value x_j by its relative frequency measure $P(x_j)$, we have used a relative weighting of each value to produce the arithmetic mean; so, computationally, we need only sum these products $x_j \cdot P(x_j)$ to

produce our arithmetic mean. In grouped data of this relative frequency form, we can find the mean by the above process, as described symbolically by the given formula:

Sample Mean from a Relative Frequency Distribution

$$\bar{x} = \sum [x_j \cdot P(x_j)]$$

Once again, if the data in our table had been population data, then we would still perform the same calculation work using the same reasoning:

Population Mean from a Relative Frequency Distribution

$$\mu = \sum [x_j \cdot P(x_j)]$$

? Text Exercise 2.8.2

Consider the Quiz 1 data from section 2.6, this time given in the relative frequency table format below. Determine the mean from the grouped format and compare it with the earlier results.

Table 2.8.7: Grouped Relative Frequency Distribution of Quiz 1 Data

Quiz Scores	Relative Frequency
5	10%
6	30%
7	25%
8	20%
9	15%

Answer

To find the mean from this relative frequency table, we follow these steps as established in the discussion above:

1. Sum the relative frequency column $P(x_j)$ to check that 100% of the data is accounted for in the table.
2. Compute the column of values $x_j \cdot P(x_j)$ to weight each of the various quiz scores with their relative frequency of occurrence.
3. Sum the column of $x_j \cdot P(x_j)$ values to produce mean of the data values.

Table 2.8.8 Computation of arithmetic mean using relative frequencies from Table 2.8.7

x_j	$P(x_j)$	$x_j \cdot P(x_j)$
5	10%	0.5
6	30%	1.8
7	25%	1.75
8	20%	1.60
9	15	1.35
Totals:	100%	$\mu = 7.00$

Again, this is the exact arithmetic mean value computed previously when working with the raw data set or the grouped frequency table set.

We extend these ideas one more step with the concept of "weighted" averages.

Weighted Mean Measures

Sometimes, data values are assigned different weights; for example, course averages are often determined through a "weighting" of the various assessment values. This weighting is usually given as a percentage but can be shown in any chosen relative form (such as a "2" weight for those values that carry twice the weight of any values assigned a "1" weight). As such, we can see how the weights play the same role as the frequency or relative frequency values in the above discussion.

As an example, suppose a school, as is commonly done, uses a four-point scale (A = 4 points, B = 3 points, C= 2 points, D = 1 point, and U= 0 points) to determine grade point average (GPA) weighted by the number of credit hours for the class. A randomly chosen student's recent letter grades awarded and number of credits in eight courses were as follows: A with 3 credits, U with 2 credits, C with 4 credits, A with 5 credits, B with 3 credits, B with 3 credits, C with 5 credits, and D with 3 credits. We organize this information in table 2.8.3 to determine this student's GPA.

Table 2.8.9: Grouped Frequency Distribution

Letter Grade	Point Value	Credit Hours (Weight)
A	4	8
B	3	6
C	2	9
D	1	3
U	0	2

Again, we use the above ideas to compute the GPA, a weighted mean.

Table 2.8.10: Computation of GPA as a weighted mean

Letter Grade	Point Value (x_j)	Credit Hours (w_j)	$x_j \cdot w_j$
A	4	8	32
B	3	6	18
C	2	9	18
D	1	3	3
U	0	2	0
	Totals:	$\sum w_j = 28$	$\sum (x_j \cdot w_j) = 71$
		Weighted Mean:	$\frac{\sum (x_j \cdot w_j)}{\sum w_j} \approx 2.5357$

This student had a GPA of 2.5357 for those courses. In data that carries varied weights, we can determine the mean as described symbolically.

Sample Mean from Weighted Data

$$\frac{\sum (x_j \cdot w_j)}{\sum w_j}$$

As we have seen, we do not always need "raw" data, especially with huge data sets, to formulate many of our descriptive statistics for the data. Grouped data conserves space required to represent data and can often be used to produce many summary statistic measures with minor adjustments to our computational thinking. However, we must know if we have any "loss" in the data representation due to the grouping. All the above data sets were discrete, and each grouping was done on single values, not over interval values. When we group data over interval values, we lose some information in the data. The following optional subsection examines this issue of continuous data.

Note: Level of measurement and careful consideration of results

An astute reader will notice that the four-point grading scale, common in many academic institutions, takes values on an ordinal scale; the arithmetic differences in values do not provide any information other than the underlying ordering of letter grades. If one student earns a 99% while a second student earns 90.1%, both students would be awarded the same letter grade of an A, despite having achieved different levels of performance in the course.

When we look at a semester's average GPA, as we did above, how are we to interpret two students in the same courses having the same average? Just like with the racing example at the end of section 1.6, we cannot say that they performed (earned points), on average, the same. One student could have outperformed the other student on all assessments in each class, yet still be awarded the same letter grades in each class thus earning an equivalent GPA. All we can say is that the students earned, on average, the same letter grades.

Text Exercise 2.8.3

Consider two physics majors, Aaron and Elise, who took Engineering Physics I (five credit hours), Calculus I (five credit hours), and Elements of Statistics (three credit hours) last semester. Aaron earned 85%, 96%, and 98%, respectively, and Elise earned 90%, 91%, and 98%, respectively.

1. Convert each student's semester grades to the four-point grading scale and then compute the weighted average using the number of credit hours as the weight. This is the standard way four-point scale averages are computed.

Answer

Aaron would receive a 3 for his physics course, and then 4 for each of his math courses. Since physics and calculus were five credit hour courses, those two grades will be weighted by 5, and statistics will be weighted by 3. We thus have the following computation.

$$\text{GPA}_{\text{Aaron}} = \frac{3\cdot 5+4\cdot 5+4\cdot 3}{5+5+3} = \frac{15+20+12}{13} = \frac{47}{13} \approx 3.6154$$

Elise earned an A in each course thus earning 4 in each of her courses. Since physics and calculus were five credit hour courses, those two grades will be weighted by 5, and statistics will be weighted by 3. We thus have the following computation.

$$\text{GPA}_{\text{Elise}} = \frac{4\cdot 5+4\cdot 5+4\cdot 3}{5+5+3} = \frac{20+20+12}{13} = \frac{52}{13} = 4$$

We thus have that Aaron earned a $3.6154$and Elise earned a 4.0 last semester.

2. Compute each student's weighted average percentage using the number of credit hours as the weight and then convert the averages to the four-point scale. This is a nonstandard way to compute four-point scale averages.

Answer

We compute the weighted averages similarly.

$$\text{GPA}_{\text{Aaron}} = \frac{85\cdot 5+96\cdot 5+98\cdot 3}{5+5+3} = \frac{425+480+294}{13} = \frac{1199}{13} \approx 92.2308\%$$

$$\text{GPA}_{\text{Elise}} = \frac{90\cdot 5+91\cdot 5+98\cdot 3}{5+5+3} = \frac{450+455+294}{13} = \frac{1199}{13} \approx 92.2308\%$$

In converting the two weighted averages to the four-point scale, both Aaron and Elise would receive a 4.0 for the semester. Despite having the same average percentages, the standard way of computation distinguishes between a 4.0 student, Elise, and Aaron, a student who did not get straight A's. There is only one way to get a 4.0. There are many ways to get a lower GPA. The four-point scale emphasizes the distinction between straight A students and everyone else.

2.8.1: Measures of Median and Mean - Grouped Data Loss of Information - Optional Material

Learning Objectives

- Consider the loss of information with grouped data
- Discuss class approximations
- Develop methods to approximate the median and mean from grouped data

Section 2.8.1 Excel File (contains all of the data sets for this section)

Central Measures on Grouped Data with Loss Of Information

What if we have data grouped over intervals instead of discrete single value groups as previously? In this case, we have lost some information about the specific data values and are only able to roughly estimate the mean and median measures of the distribution. Below is a frequency/relative frequency table, Table 2.8.1.2 based on data given by Florence Nightingale in her text *Notes on Nursing* (downloaded here). The text listed the ages of a large sample of non-domestic servant nurses within Great Britain in 1851 in a grouped data interval format. We will assume that Ms. Nightingale collected the data in a way such that if, for example, someone was in their 29^{th} year of age (such as 29.875 years old), the data was reported as a 29 and not rounded up to 30...a common convention in reporting of ages for individuals. We have added the interval notation representation of the continuous variable of age per that convention to the table.

Table 2.8.1.1: Grouped frequency distribution

Age Intervals (years)	Interval Notation (years)	Frequency	Relative Frequency
$20-30$	$[20, 30)$	$1,441$	$\frac{1,441}{25,466} \approx 0.0566 = 5.66\%$
$30-39$	$[30, 40)$	$2,477$	$\frac{2,477}{25,466} \approx 0.0973 = 9.73\%$
$40-49$	$[40, 50)$	$4,971$	$0.1952 = 19.52\%$
$50-59$	$[50, 60)$	$7,438$	$0.2921 = 29.21\%$
$60-69$	$[60, 70)$	$6,367$	$0.2500 = 25.00\%$
$70-79$	$[70, 80)$	$2,314$	$0.0909 = 9.09\%$
$80+$	$[80, \text{above})$	458	$0.0180 = 1.80\%$
Totals:		$25,466$	$1.0000 = 100\%$

Notice we do not know how many 20 year old nurses there were in the data set, nor do we know how many 27 year old nurses there were. We only know that there were $1,441$ nurses reporting ages of $20-29$. This means that we cannot know what the actual data values were in the original data set; we have lost specific information about the original data set.

We can, however, approximate descriptive statistics based on this grouped data. We will proceed as in the discrete case above, except we will use the midpoint value of each interval as our best approximation single measure for all values within the interval. For example, we will assume that all 1441 people in their twenties are exactly 25 years old, the midpoint of that interval. This is a drastic assumption in some sense, but with the loss of information on specific age measures in each interval, this is a reasonable way to approximate our measures. We will also use 85 as our value for the last class interval of $[80, \text{above})$ even though the midpoint value may be larger if more was known about the actual data. It is reasonable to believe that in 1851 most nurses above the age of 80 were likely closer to the 80 value than the 85 value; but this is an assumption we are making and must be disclosed.

These assumptions, across all the intervals, will only give us estimates of the actual true mean and median measures of center. So, for the median, we begin to accumulate our relative frequencies until we know where the 50^{th} percentile measure lies. Since $5.66\% + 9.73\% + 19.52\% = 34.91\%$, which is less than 50%, and $5.66\% + 9.73\% + 19.52\% + 29.21\% = 64.12\%$, which is

greater than 50%, we know the 50^{th} percentile location is within the interval $50-59$. Thus our estimate for the median would be 55 years old.

To estimate the arithmetic mean, we can use the midpoint of each interval as the data value associated with each of the relative frequency measures and complete our computation work as in the discrete case.

Table 2.8.1.2 Computation of mean using data from Table 2.8.1.1

Age Intervals (years)	Midpoint (m_j) (years)	$P(m_j)$	$m_j \cdot P(m_j)$
$20-29$	25	$0.0566 = 5.66\%$	$25 \cdot 0.0566 = 1.4150$
$30-39$	35	$0.0973 = 9.73\%$	$35 \cdot 0.0973 = 3.4055$
$40-49$	45	$0.1952 = 19.52\%$	$45 \cdot 0.1952 = 8.7840$
$50-59$	55	$0.2921 = 29.21\%$	16.0655
$60-69$	65	$0.2500 = 25.00\%$	16.2500
$70-79$	75	$0.0909 = 9.09\%$	6.8175
$80+$	85	$0.0180 = 1.80\%$	1.5300
	Totals:	$1.0000 = 100\%$	$\sum (m_j \cdot P(x_j)) \approx 54.2675$

So, we would estimate the mean age of all these sampled non-domestic servant nurses in Great Britain to be about 54.3 years old. In examining the relative frequency measures as tied to the age intervals, this value makes reasonable sense as the "balance point" of the distribution of the ages. So, in grouped data within intervals, we can estimate the mean by the same overall process, described symbolically by the given formula with the use of each interval's midpoint represented by m_j :

Mean from an Interval-Grouped Distribution

$$\bar{x} \approx \frac{\sum (m_j \cdot f_j)}{\sum f_j} = \sum (m_j \cdot P(m_j)) \text{ when working with interval grouped sample data}$$

$$\mu \approx \frac{\sum (m_j \cdot f_j)}{\sum f_j} = \sum (m_j \cdot P(m_j)) \text{ when working with interval grouped population data}$$

In summary, we have seen how we can still determine estimates for the median and mean measurement when given interval grouped data.

? Text Exercise 2.8.1.1

A bakery has been keeping records on the shelf-life of its best selling cinnamon rolls package. The bakery has sent the following frequency table asking for the median and mean measures of the data. Find reasonable estimates of the mean and the median values of the data.

Table 2.8.1.3 Grouped frequency distribution for shelf-life data

Shelf-life (days)	Frequency
$[3, 8)$	3
$[8, 13)$	19
$[13, 18)$	43
$[18, 23)$	21
$[23, 28)$	16

Shelf-life (days)	Frequency
$[3, 8)$	3
$[8, 13)$	19
$[13, 18)$	43
$[18, 23)$	21
$[23, 28)$	16
$[28, 33)$	2

Answer

We proceed by extending our table to include a column of midpoint values and to compute relative frequency measures. Do note we could also use straight frequency as a weighting measure, but choose to use the relative frequency approach instead.

Table 2.8.1.4 Preparatory computations using data from Table 2.8.1.3

Shelf-life (days)	Midpoint (m_j) (days)	Frequency	Relative Frequency (P \left(m_j \right) \)
$[3, 8)$	$\frac{3+8}{2} = 5.5$	3	$\frac{3}{104} \approx 0.0288$
$[8, 13)$	$\frac{8+13}{2} = 10.5$	19	$\frac{19}{104} \approx 0.1827$
$[13, 18)$	15.5	43	0.4135
$[18, 23)$	20.5	21	0.2019
$[23, 28)$	25.5	16	0.1538
$[28, 33)$	30.5	2	0.0192
	Totals:	104	1.0000

To estimate the median, we again focus on our relative frequency measures to get a "location". We notice that $2.88\% + 18.27\% = 21.15\%$, which is less than 50%, and $2.88\% + 18.27\% + 41.35\% = 62.50\%$, which is greater than 50%. The 50^{th} percentile location is within the interval $[13, 18)$. Thus our estimate for the median shelf-life of the packages of cinnamon rolls by this bakery would be 15.5 days.

Next, we weight each midpoint value by its corresponding relative frequency measure, before summing to produce our mean measure.

Table 2.8.1.5 Computation of mean shelf-life

Shelf-life (days)	Midpoint (m_j) (days)	$P(m_j)$	$m_j \cdot P(m_j)$
$[3, 8)$	5.5	0.0288	$5.5 \cdot 0.0288 \approx 0.1587$
$[8, 13)$	10.5	0.1827	$10.5 \cdot 0.1827 \approx 1.9183$
$[13, 18)$	15.5	0.4135	6.4087
$[18, 23)$	20.5	0.2019	4.1304
$[23, 28)$	25.5	0.1538	3.9231
$[28, 33)$	30.5	0.0192	0.5865
	Totals:	1.0000	17.1346

So, our estimate for the mean shelf-life of the packages of cinnamon rolls by this bakery would be about 17.1 days.

In summary, we have seen how we can determine estimates for the median and mean measurement when given interval-grouped data, but also heed the warning that these are just rough estimates and that we must not considered our results as the actual measures for the data that was originally collected.

2.9: Measures of Variance and Standard Deviation on Grouped Data

Learning Objectives

- Determine range, variance, and standard deviation of grouped discrete data
- Determine range, variance, and standard deviation of grouped relative frequency data

Section 2.9 Excel File (contains all of the data sets for this section)

Introduction to Measures of Spread on Grouped Data

As mentioned in the previous section, there are times when we may be given data in a summarized frequency distribution format instead of a collection of "raw" data. We now examine finding descriptive statistic measures of dispersion in grouped data: range, variance, and standard deviation. If the data is grouped over intervals, we can only estimate such measures since the grouping action has caused us to lose some data information. However, if the data is grouped into single-value classes, we can usually produce the same spread measures as if we had the raw data.

Range, Variance, and Standard Deviation of Grouped (non-interval) Data in a Frequency Table

Look at the frequency table from Text Exercise 2.8.1 regarding quiz scores of twenty students:

Table 2.9.1: Grouped (non-interval) Frequency Distribution of Quiz 1 Data

Quiz Scores	Frequency
5	2
6	6
7	5
8	4
9	3

With such a table, we can easily find the range with no new ideas needed. The table shows the minimum data value is 5 and the maximum is 9. Because $9-5=4$, the range is a score difference of 4.

Next, assuming the data is population data, we examine the variance and standard deviation. Recall that variance is the average of all the various squared deviations of the individual data values from the mean of the data. In the previous section, we found the mean of this distribution to be $\mu=\frac{\sum(x_j\cdot f_j)}{\sum f_j}=\frac{140}{20}=7$. In a new column, we can construct the deviations from the mean and square those deviations as a first step. It becomes imperative that we not get tied up with all the messy numbers as we move through our work but keep our focus on what we are measuring:

Table 2.9.2: Computation of squared deviations from the mean

x_j	f_j	$(x_j-\mu)^2$
5	2	$(5-7)^2=4$
6	6	$(6-7)^2=1$
7	5	$(7-7)^2=0$
8	4	$(8-7)^2=1$
9	3	$(9-7)^2=4$
Totals:	$\sum f=20$	

Before we can average these squared deviation measures, we must remember that some squared deviations occur more frequently (as given by the frequency column) than others. For example, the frequency column shows that many more squared deviations are tied to the data values of 6 than those of the squared deviations associated with the data value of 9. We must "weight" these squared deviations by the frequency of their occurrence to account for the various twenty data values. In another new column, we form the products of $(x_j-\mu)^2\cdot f_j$ to accomplish this, and then proceed to "average" those weighted squared deviations:

Table 2.9.3: Computation of the variance

x_j	f_j	$(x_j-\mu)^2$	$(x_j-\mu)^2\cdot f_j$
5	2	4	$4\cdot 2=8$
6	6	1	$1\cdot 6=6$
7	5	0	$0\cdot 5=0$

x_j	f_j	$(x_j-\mu)^2$	$(x_j-\mu)^2 \cdot f_j$
8	4	1	$1 \cdot 4 = 4$
9	3	4	$4 \cdot 3 = 12$
Totals:	$\sum f_j = 20$		$\sum (x_j-\mu)^2 \cdot f_j = 30$
		Variance:	$\sigma^2 = \frac{\sum\left[(x_j-\mu)^2 \cdot f_j\right]}{\sum f_j} = \frac{30}{20} = 1.5$

As shown in the last two rows above, with our weighted squared deviations determined, we can find a meaningful average by summing our squared deviations and dividing by the number of data values involved. We have a variance measure of σ^2\(=1.5\) in the bottom right table cell. We can find the standard deviation by taking the square root of our variance: $\sigma = \sqrt{\sigma^2} = \sqrt{1.5} \approx 1.2247$. We note that our calculation work was for population data. If the table referenced sample data, we would have divided by $\sum(f_j) - 1$ instead on the last computation:

Variance and Standard Deviation from a Frequency Distribution

$$\text{Variance for sample data: } s^2 = \frac{\sum\left[(x_j-\bar{x})^2 \cdot f_j\right]}{\sum(f_j) - 1}$$

$$\text{Standard Deviation for sample data: } s = \sqrt{\frac{\sum\left[(x_j-\bar{x})^2 \cdot f_j\right]}{\sum(f_j) - 1}} = \sqrt{s^2}$$

$$\text{Variance for population data: } \sigma^2 = \frac{\sum\left[(x_j-\mu)^2 \cdot f_j\right]}{\sum f_j}$$

$$\text{Standard Deviation for population data: } \sigma = \sqrt{\frac{\sum\left[(x_j-\mu)^2 \cdot f_j\right]}{\sum f_j}} = \sqrt{\sigma^2}$$

As mentioned, working in a spreadsheet will often make the computation quicker and easier, especially when working with such columns of information.

? Text Exercise 2.9.1

The frequency table from Section 2.1 shows thirty student scores (discrete 10-point scale) for an assignment.

Table 2.9.4: Student quiz score data

Student Score	Frequency
3	1
4	1
5	3
6	5
7	5
8	7
9	5
10	3

Find the range, variance, and standard deviation of this grouped (non-interval) data. Assume the data is a sample from a larger population.

Answer

The table shows that the minimum data value is 3 and the maximum data value is 10. Because $10 - 3 = 7$, the range is a score difference of 7.

Next, we determine the variance by averaging the squared deviations from the mean, once weighted by the frequency counts. Since this is sample data, we also remember to divide by one less in the averaging step. In the previous section, we found the mean of this distribution to be $\bar{x} = \frac{\sum(x_j \cdot f_j)}{\sum f_j} = \frac{218}{30} \approx 7.2667$. We can construct the deviations from the mean and square those as the first step. Again, it becomes imperative that we not get distracted by the messy decimals we compute in our work. Keep focusing on what we measure in the computational work (squared deviations from the mean to be averaged) and calculate it accurately.

Table 2.9.5 Computation of square deviations from the mean

x_j	f_j	$(x_j - \mu)^2$
3	1	$(3 - 7.2667)^2 \approx 18.2044$
4	1	$(4 - 7.2667)^2 \approx 10.6711$
5	3	$(5 - 7.2667)^2 \approx 5.1378$
6	5	$(6 - 7.2667)^2 \approx 1.6044$
7	5	$(7 - 7.2667)^2 \approx 0.0711$
8	7	$(8 - 7.2667)^2 \approx 0.5378$
9	5	$(9 - 7.2667)^2 \approx 3.0044$
10	3	$(10 - 7.2667)^2 \approx 7.4711$
Totals:	$\sum f_j = 30$	

Next, we weight our various squared deviations by their frequency of occurrence, forming the products of $(x_j - \mu)^2 \cdot f_j$ to accomplish this:

Table 2.9.6 Computation of variance

x_j	f_j	$(x_j$
3	1	18.
4	1	10.
5	3	5.1
6	5	1.6
7	5	0.7
8	7	0.5
9	5	3.0
10	3	7.4
Totals:	$\sum f_j = 30$	
		V a r i a n c e : $s =$

Again, our last row of the table shows the "averaging" of those weighted squared variations when working with sample data. If the data were given as population data, we would have divided by the sum of the frequencies instead of one less than that sum. This would have resulted in a slightly different value of $\sigma^2 \approx 3.1289$. When calculating, we must vigilantly remember the difference between sample and population variance. These data measures are distinct and continue to emphasize one, among other, reasons for knowing if the data is sample data or population data. As shown by the bottom right measure in the table, this grouped sample data has a variance measure of $s^2 \approx 3.2368$. The sample standard deviation measure is: $s = \sqrt{s^2} \approx \sqrt{3.2368} \approx 1.7991$. We can compare this result with the calculated variance for the ungrouped data set to see that we have produced the same measure.

Range, Variance, and Standard Deviation of Grouped (non-interval) Data in a Relative Frequency Table

What must we do to measure the variation of the data if, instead of a frequency distribution table, we have a relative frequency distribution of population data? The approach presented here only produces valid measures in population data since the relative frequency measures do not always disclose the sample size. Recall the sample size is essential for the computation of the sample variance as we must use $n - 1$ in our averaging step. We start with the same example of twenty quiz scores.

Table 2.9.6: Grouped Relative Frequency Distribution of Quiz 1 Data

Quiz Scores	**Relative Frequency $P(x_j)$**

Quiz Scores	Relative Frequency $P(x_j)$
5	$\frac{2}{20} = 10\%$
6	$\frac{2}{20} = 30\%$
7	$\frac{5}{20} = 25\%$
8	$\frac{4}{20} = 20\%$
9	$\frac{3}{20} = 15\%$
Totals:	$\sum P(x_j) = 100\%$

Again, we can easily find the range--the table shows the minimum data value is 5 and the maximum data value is 9, leading us to a range of $9 - 5 = 4$.

Next, we use the same ideas as above to determine the variance and standard deviation of the data from this table. We need to find the average squared deviations of the data values from the mean. In Section 2.8, we found the mean of this relative frequency distribution to be $\mu = \sum (x_j \cdot P(x_j)) = 7$. We will add a column to our table to create the squared deviations from the mean.

Table $2.9.7$: Computation of the squared deviations from the mean

x_j	$P(x_j)$	$(x_j - \mu)^2$
5	$\frac{2}{20} = 10\%$	$(5-7)^2 = 4$
6	$\frac{2}{20} = 30\%$	$(6-7)^2 = 1$
7	$\frac{5}{20} = 25\%$	$(7-7)^2 = 0$
8	$\frac{4}{20} = 20\%$	$(8-7)^2 = 1$
9	$\frac{3}{20} = 15\%$	$(9-7)^2 = 4$
Totals:	$\sum P(x_j) = 100\%$	

Before we can average these squared deviation measures, we must weight these squared deviations by the relative frequency of occurrence to account for the fact that some data values, such as the 6, occur with different relative frequency than others, such as the 9. To do so, we form a new column for the products of $(x_j - \mu)^2 \cdot P(x_j)$ and then proceed to "average" those weighted squared deviations.

Table $2.9.8$: Computation of the variance

x_j	$P(x_j)$	$(x_j - \mu)^2$	$(x_j - \mu)^2 \cdot P(x_j)$
5	$\frac{2}{20} = 10\%$	4	$4 \cdot 0.10 = 0.40$
6	$\frac{2}{20} = 30\%$	1	$1 \cdot 0.30 = 0.30$
7	$\frac{5}{20} = 25\%$	0	$0 \cdot 0.25 = 0.00$
8	$\frac{4}{20} = 20\%$	1	$1 \cdot 0.20 = 0.20$
9	$\frac{3}{20} = 15\%$	4	$4 \cdot 0.15 = 0.60$
Totals:	$\sum P(x_j) = 100\%$		$\sum \left((x_j - \mu)^2 \cdot P(x_j)\right) = 1.50$
		Variance:	$\sigma^2 = \frac{\sum\left[(x_j-\mu)^2 \cdot P(x_j)\right]}{\sum P(x_j)} = \frac{1.50}{1} = 1.5$

As shown in the last two rows, with our weighted squared deviations determined, we can find the average by summing our squared deviations and then dividing by the total weighting of the relative frequency measures, which will always be the value $1.00 = 100\%$. We again get the same measure of variance, 1.5, as we did earlier when the data was in frequency table form.

Variance and Standard Deviation from a Relative Frequency Distribution

$$\text{Variance for population data: } \sigma^2 = \sum \left[(x_j - \mu)^2 \cdot P(x_j)\right]$$

$$\text{Standard Deviation for population data: } \sigma = \sqrt{\sum \left[(x_j - \mu)^2 \cdot P(x_j)\right]} = \sqrt{\sigma^2}$$

We should note that the computation work here is simpler when the grouped data is given in relative frequency rather than just frequency format. This is one reason we often look at data in relative frequency form.

Text Exercise 2.9.2

We take the frequency table shown below from Section 2.1 regarding thirty student scores (discrete 10-point scale) for an assignment but with the distribution given in relative frequency format instead. This time, we assume the data is population data—our focus is only on these thirty students and not some larger group.

Table 2.9.9: Relative frequency distribution of student scores

x_j	$P(x_j)$
3	$\frac{1}{30} \approx 0.0333 = 3.33\%$
4	$\frac{1}{30} \approx 0.0333 = 3.33\%$
5	$\frac{3}{30} = 0.1000 = 10.00\%$
6	$0.1667 = 16.67\%$
7	$0.1667 = 16.67\%$
8	$0.2333 = 23.33\%$
9	$0.1667 = 16.67\%$
10	$0.1000 = 10.00\%$
Totals:	$\sum P(x_j) = 1.0000 = 100\%$

Find this grouped population data's range, variance, and standard deviation.

Answer

The given table clearly shows that the minimum data value is 3 and the maximum is 10, thus a range measure of 7.

For variance, we must first find the squares on the deviations from the mean, as shown in the added column below. Recall that, in Section 2.8, we found the mean of the relative frequency distribution as $\mu = \sum x_j \cdot P(x_j) \approx 7.2667$. We will use this value in our computation work.

Table 2.9.10 Computation of squared deviations from the mean

x_j	$P(x_j)$	$(x_j - \mu)^2$
3	$0.0333 = 3.33\%$	$(3 - 7.2667)^2 \approx 18.2044$
4	$0.0333 = 3.33\%$	$(4 - 7.2667)^2 \approx 10.6711$
5	$0.1000 = 10.00\%$	$(5 - 7.2667)^2 \approx 5.1378$
6	$0.1667 = 16.67\%$	$(6 - 7.2667)^2 \approx 1.6044$
7	$0.1667 = 16.67\%$	$(7 - 7.2667)^2 \approx 0.0711$
8	$0.2333 = 23.33\%$	$(8 - 7.2667)^2 \approx 0.5378$
9	$0.1667 = 16.67\%$	$(9 - 7.2667)^2 \approx 3.0044$
10	$0.1000 = 10.00\%$	$(10 - 7.2667)^2 \approx 7.4711$
Totals:	$\sum P(x_j) = 1.0000 = 100\%$	

Now we weight, through multiplication, our various squared deviations by their relative frequency of occurrence. This forms the products $(x_j - \mu)^2 \cdot P(x_j)$ in our next added column.

Table 2.9.11: Computation of the variance

x_j	$P(x_j)$	$(x_j - \mu)^2$	$(x_j - \mu)^2 \cdot P(x_j)$
3	$0.0333 = 3.33\%$	18.2044	$18.2044 \cdot 0.0333 \approx 0.6068$
4	$0.0333 = 3.33\%$	10.6711	$10.6711 \cdot 0.0333 \approx 0.3557$
5	$0.1000 = 10.00\%$	5.1378	$5.1378 \cdot 0.1000 \approx 0.5138$
6	$0.1667 = 16.67\%$	1.6044	$1.6044 \cdot 0.1667 \approx 0.2674$
7	$0.1667 = 16.67\%$	0.0711	$0.0711 \cdot 0.1667 \approx 0.0119$
8	$0.2333 = 23.33\%$	0.5378	$0.5378 \cdot 0.2333 \approx 0.1255$
9	$0.1667 = 16.67\%$	3.0044	$3.0044 \cdot 0.1667 \approx 0.5007$

10	$0.1000 = 10.00\%$	7.4711	$7.4711 \cdot 0.1000 \approx 0.7471$
Totals:	$\sum P(x_j) = 1.0000 = 100\%$		
		Variance:	$\sigma^2 = \sum \left[(x_j - \mu)^2 \cdot P(x_j) \right] \approx 3.1289$

Our last row illustrates the production of the variance as the average of the weighted squared deviations. The original table of data has a population variance of $\sigma^2 \approx 3.1289$and population standard deviation of $\sigma = \sqrt{\sigma^2} \approx \sqrt{3.1289} \approx 1.7689$.

We have computed the three desired measures of variation in the data and grouped them within a relative frequency table.

We can compute population variance and standard deviation even when given data in a relative frequency table. However, we cannot do so for sample data given only in a relative frequency table.

Section Summary

This section has demonstrated that we can often compute range, variance, and standard deviation even after the data has been grouped into a frequency table or a relative frequency table. We also remind ourselves that the formulas developed in this section came from the meaning of each measure. We do not memorize the formulas; we recall what each measure means and how and why we performed the computations. This section demonstrated how we can adjust our process to produce the same measures, but it also showed that, due to the "column" computation work, the use of a spreadsheet makes the process much easier.

The following optional section explores how we might estimate the range, variance, and standard deviation if our data has been grouped into interval classes. The ideas are similar, but we can only roughly estimate such measures because we have lost individual data representation (a loss of information about the data).

2.9.1: Measures of Variance and Standard Deviation - Loss of Information - Optional Material

Learning Objectives

- Consider the loss of information with grouped data
- Discuss class approximations
- Develop methods to approximate the range, variance, and standard deviation from grouped data

Section 2.9.1 Excel File (contains all of the data sets for this section)

Dispersion Measures on Interval-Grouped Data with Loss Of Information

What if we have data grouped over intervals instead of discrete single value groups as previously in Section 2.9? As with our measures of central tendency, we have lost some information about the specific data values and are only able to roughly estimate our dispersion measures of range, variance, and standard deviation measures. We examine common approximation methods below; however, it should be known that the methods shown below are not unique as other similar but different choices can be made in what values are used from the class intervals.

The following table is again the frequency/relative frequency table based on data given by Florence Nightingale in her text *Notes on Nursing* (downloaded here). We will again assume that Ms. Nightingale collected the data in a way such that if, for example, someone was in their 29^{th} year of age (such as 29.875years old), the data was reported as a 29. We will also take this information as a representation of population data in our following computation work.

Table 2.9.1.1: Grouped Frequency Distribution

Age Intervals (years)	Interval Notation (years)	Frequency	Relative Frequency
$20-30$	$[20, 30)$	$1,441$	$\frac{1,441}{25,466} \approx 0.0566 = 5.66\%$
$30-39$	$[30, 40)$	$2,477$	$\frac{2,477}{25,466} \approx 0.0973 = 9.73\%$
$40-49$	$[40, 50)$	$4,971$	$0.1952 = 19.52\%$
$50-59$	$[50, 60)$	$7,438$	$0.2921 = 29.21\%$
$60-69$	$[60, 70)$	$6,367$	$0.2500 = 25.00\%$
$70-79$	$[70, 80)$	$2,314$	$0.0909 = 9.09\%$
$80+$	$[80, \text{above})$	458	$0.0180 = 1.80\%$
Totals:		$25,466$	$1.0000 = 100\%$

We recall that we do not know, for example, how many 20 year old nurses there were in the data set, nor do we know how many 27 year old there were. We only know that there were $1,441$ nurses reporting ages of $20-29$. This clearly implies that we cannot know what the actual data values were in the original data set; we have lost specific information about the original data set.

We will similarly approximate our measures of dispersions based on this grouped data by use of the midpoint value of each interval as our best approximation single measure for all values within the interval. So, again we will assume that all $1,441$ people in their twenties are 25 years old, the midpoint of that interval. This is a drastic assumption in some sense, but with the loss of information on specific age measures in each interval, this is a reasonable way to approximate our dispersion measures just as with our central tendency measures. We will also again use 85 as our value for the last class interval of $[80, \text{above})$ even though the midpoint value could be larger if more was known about the actual data. This is also pointing to why it is not a general best practice when building frequency tables of data to use "and above" or "and below" within the last and first class interval descriptions; doing so provides even greater loss of key information about the data set.

We might choose to estimate the range measure using the largest and smallest midpoint values. That is, we estimate the range to be approximately $85-25 = 60$ years. We note that we consider this a very rough estimate and major decisions should not be based on this estimate. Likely the range measure is larger, but again due to the loss of information when the data was grouped, we can't know for sure. We also note that others might estimate the range from such grouped data differently (such as the highest class' upper limit value minus the lowest class' lower limit value.)

Now for the variance estimate. First, we also recall from the optional Section 2.8.1, that we computed a mean estimate value of $\mu \approx 54.2675$ years old. Variance, as the average of squared deviations from the mean, then leads to our producing the squared deviations column and weighting of those squared deviations by the relative frequency measures (we choose to use the relative frequency versus frequency approach in our work). Again, we use the midpoint of each interval as the data value to estimate deviation from the mean measures, and complete our work as in the non-interval grouped data approach of Section 2.9.

Table 2.9.1.2 Computation of variance

Age Intervals (years)	Midpoint (m_j) (years)	$P(m_j)$	$(m_j-\mu)^2 \cdot P(m_j)$
$20-29$	25	$0.0566 = 5.66\%$	$(25-54.2675)^2 \cdot 0.0566 \approx 48.4828$
$30-39$	35	$0.0973 = 9.73\%$	$(35-54.2675)^2 \cdot 0.0973 \approx 36.1213$
$40-49$	45	$0.1952 = 19.52\%$	$(45-54.2675)^2 \cdot 0.1952 \approx 16.7651$
$50-59$	55	$0.2921 = 29.21\%$	0.1567
$60-69$	65	$0.2500 = 25.00\%$	28.7966
$70-79$	75	$0.0909 = 9.09\%$	39.0721
$80+$	85	$0.0180 = 1.80\%$	17.0008
	Totals:	$1.0000 = 100\%$	$\sum\left[(m_j-\mu)^2 \cdot P(m_j)\right] \approx 186.3954$

So, we would estimate the variance of all this given population of non-domestic servant nurses in Great Britain to be about $\sigma^2 \approx 186.3954 \text{years}^2$, and hence the standard deviation to be about $\sigma \approx \sqrt{186.3954} \approx 13.6527$ years. So, with interval-grouped data, we can estimate the variance and standard deviation by the same overall process, described symbolically by the given formulas with the use of each interval's midpoint represented by m_j :

Variance from an Interval-Grouped Distribution

$$s^2 \approx \frac{\sum\left[(m_j-\bar{x})^2 \cdot f_j\right]}{\sum f_j - 1} \text{ ; when working with frequency interval-grouped sample data}$$

$$\sigma^2 \approx \frac{\sum\left[(m_j-\mu)^2 \cdot f_j\right]}{\sum f_j} = \sum\left(m_j \cdot P(x_j)\right) \text{ when working with frequency or relative frequency interval-grouped population data}$$

Standard Deviation from an Interval-Grouped Distribution

$s \approx \sqrt{s^2}$ when working with interval-grouped sample data

$\sigma^2 \approx \sqrt{\sigma^2}$ when working with interval-grouped sample data

? Text Exercise 2.9.1.1

A bakery has been keeping records on the shelf-life of its best selling cinnamon rolls package. The bakery has sent the following frequency table asking for the median and mean measures of the data. Assuming this is sample data, find reasonable estimates of the range, variance, and standard deviation values of the data.

Table 2.9.1.3 Grouped frequency distribution for shelf-life data

Shelf-life (days)	Frequency
$[3, 8)$	3
$[8, 13)$	19
$[13, 18)$	43

Shelf-life (days)	Frequency
$[3, 8)$	3
$[8, 13)$	19
$[13, 18)$	43
$[18, 23)$	21
$[23, 28)$	16
$[28, 33)$	2

Answer

We again proceed by extending our table to include a column of midpoint values. Since this is sample data, we keep with frequency versus relative frequency measures in order to not lose sample size information.

Table 2.9.1.4 Preparatory computations using data from Table 2.9.1.3

Shelf-life (days)	Midpoint (m_j) (days)	Frequency f_j
$[3, 8)$	$\frac{3+8}{2} = 5.5$	3
$[8, 13)$	$\frac{8+13}{2} = 10.5$	19
$[13, 18)$	15.5	43
$[18, 23)$	20.5	21
$[23, 28)$	25.5	16
$[28, 33)$	30.5	2
	Totals:	104

First, we estimate the range to be $30.5 - 5.5 = 25$ days using our midpoint values. (As mentioned in the discussion above, one might instead choose to compute $33 - 3 = 30$ days for the range; this estimate would be considered a maximum amount the range might truly be.)

Next, we estimate the variance. In Section 2.8.1,we estimated the mean of this data to be 17.1346days. So, to estimate the sample variance, we must form the weighted squared variations from the mean column, then sum those squared variations, and finally divide by one less than the sample size to form our "average" of the squared variations for sample variance purposes.

Table 2.9.1.5 Computation of variance

Shelf-life (days)	Midpoint (m_j) (days)	f_j	$(m_j - \mu)^2 \cdot f_j$
$[3, 8)$	5.5	3	$(5.5 - 17.1346)^2 \cdot 3 \approx 406.0918$
$[8, 13)$	10.5	19	$(10.5 - 17.1346)^2 \cdot 19 \approx 836.3404$
$[13, 18)$	15.5	43	$(15.5 - 17.1346)^2 \cdot 43 \approx 114.8924$
$[18, 23)$	20.5	21	237.8443
$[23, 28)$	25.5	16	1119.6787
$[28, 33)$	30.5	2	357.2678
	Totals:	104	$s^2 \approx \frac{\sum\left[(m_j-\mu)^2 \cdot f_j\right]}{\sum(f_j)-1} \approx \frac{3072.1154}{104-1} \approx 29.$

So, our estimate for the variance on the shelf-life of the packages of cinnamon rolls by this bakery would be about $s^2 \approx 29.8264$ days^2. And thus our standard deviation estimate would be $s \approx \sqrt{29.8264} \approx 5.4614$days. So the cinnamon roll packages roughly tend to last about 17.1 ± 5.5 days.

In summary, we have now seen how we can produce rough estimates for the dispersion measurements when given interval-grouped data. We note how we are really just extending previous ideas/computations. However, we also remind ourselves that the resulting values should be used with caution in the interpretation of the dispersion of the data, not as if the values were the actual true measures of the data.

CHAPTER OVERVIEW

3: Probability

3.1: Introduction to Probability

Learning Objectives

- Learn key initial terminology about probability
- Determine the sample space of a given situation
- Recognize and use the three basic methods of determining probability measures
- Explain the importance of the *Law of Large Numbers*
- Describe the complement of an event
- Use the Complement Probability Rule for determining the probability of an event

Review and Preview

Inferential statistics seeks to make educated guesses about populations using statistics from randomly chosen samples. The usefulness of a sample statistic depends on the sample from which it was taken. We cannot guarantee that samples are representative of the population, but we can ensure that any bias present is due to random chance. What is the likelihood that our randomly chosen sample is representative? In other words, what is the probability that our sample statistic is close to the population parameter? These are fundamental questions to examine through the science of probability.

We hear probability claims daily. The weather forecast states a 30% chance of rain. The probability of a faulty product coming off a manufacturing line is 6.5%. One has a 20% probability of purely guessing an answer correctly on a single multiple-choice question. A cancer research group believes that 40% of women and 45% of men will have a diagnosis of some type of cancer during their lifetimes. (Note: This means that if we randomly select a man, we have a 45% chance of choosing someone who has had or will have a cancer diagnosis in his life. This does not mean that any specific man has a 45% chance of being diagnosed with cancer in his lifetime.)

Each of the above scenarios involves a situation in which something will happen, and an outcome will occur, but we are uncertain which outcome it will be. Will it rain, or will it not rain? Is the next product produced of high quality or not? In statistics, we refer to such situations as random experiments. We have a clear context and an idea of possible outcomes, one of which will happen, and then use probabilities to measure the likelihood of any particular outcome.

As we move through the course, we will develop an understanding of random sampling to build probabilities. We will develop the ability to measure how unusual a sample statistic is within a given situation. If the probability of a calculated sample statistic is "small," then we conclude that the outcome is unusual. The use of probability goes well beyond this application; therefore, we develop probability within a larger context.

Basic Concept of Probability

We begin by examining the meaning of the term probability. Generally, **probability** is a numerical value inclusively between 0 and 1 measuring the likelihood that a specific event will occur within a given situation. The representation of that numerical value might be in decimal form, fractional form, or percentage form, such as $0.75 = \frac{3}{4} = 75\%$. An **outcome** of a random experiment is a potential result of the experiment. The term **event** is a set of outcomes one might expect from a given random experiment. For example, if one rolls a pair of standard game dice as shown in Figure $3.1.1$, a possible event could be both dice landing with one facing up.

Figure $3.1.1$: Game Dice

Before measuring probability accurately, we must clearly describe the given situation (such as rolling a pair of standard game dice) and the event of interest (such as both dice landing with one facing up). For a given clearly described situation, the collection of all

possible outcomes is called the **sample space** or **event space**. For reasonably simple situations, one can fully describe the sample space.

Text Exercise 3.1.1

What is the sample space if one rolls a pair of standard game dice, as shown in Figure 3.1.1?

Answer

The graphic below shows the sample space, where one die outcome is red and the other in white. Notice there are thirty-six possible outcomes. We should also notice that, in general, we must be concerned with the order of the two dice. For example, we consider rolling a three on the red die, and a one on the white die a different outcome than rolling a one on the red die and a three on the white die. Why is this? Because these are two outcomes that we can distinguish. Since probability measures the degree of uncertainty, one must consider all available information. We produce a complete sample space by carefully reflecting on all possible outcomes.

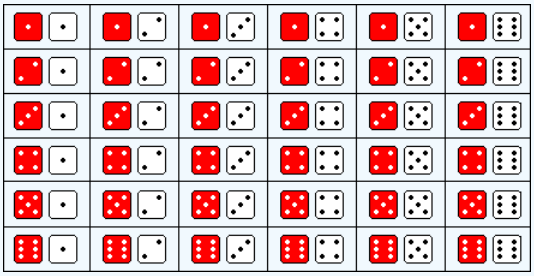


Figure 3.1.2 Sample space of rolling a pair of standard game dice

When rolling a pair of dice, the event of rolling two ones is an outcome. However, the event in which both dice land with a sum of five describes several outcomes in the sample space: 1 and 4; 2 and 3; 3 and 2; and 4 and 1.

Once we understand the sample space, we can examine the probability measures of various events within that situation. A probability measure near 1 indicates that the specific event is more likely to occur, and a probability measure near 0 indicates that the specific event is less likely to occur. We use the symbols $P(A)$ to indicate the probability of a specific event A. For example, $P(\text{rain}) = .30 = 30\%$ indicates "the probability of rain is 30%". If for some event A, we find that $P(A) = 1.00 = 100\%$, then we say this event is a **certain event**. Similarly, if for some event A, we find that $P(A) = 0.00 = 0\%$, then we say this event is an **impossible event**. For finite sample spaces, impossible events cannot occur, and certain events must occur. For infinite sample spaces, the situation is more complicated; we will discuss this further in chapter 4. For now, we are usually interested in events that are possible but not likely to occur.

Note: Unusual Events

In inferential statistics, we are often interested in outcomes that are not likely to happen; that is, they are not very probable; they are unusual. When improbable outcomes occur, either something rare happens, or we have reason to think our understanding of the situation needs to be updated, but what is the cutoff between improbable and probable, between unusual and usual? There is no fixed, agreed-upon value that will work for every situation. For this reason, we will often need to set a probability measure for considering an event improbable. Different individuals can generally feel the same event usual/likely or unusual/unlikely, depending on their personal life experiences. For example, if the probability of rain is 15%, one person may consider rain unlikely, and another may think it reasonably likely. A commonly chosen boundary measure for many statistical analysis areas is 5%. We will initially require a probability measure of 5% or less to label an event as unusual. Later, we will relax on this requirement, even leaving the choice to the consumer of our statistical measures to apply their chosen probability measure for unusual.

? Text Exercise 3.1.2

1. Explain why 150% cannot be the probability of some event.

Answer

The value 150% is larger than 100%. Any valid probability value will be between $0.00 = 0\%$ and $1.00 = 100\%$, inclusively. Knowing such will help us recognize improper probabilities.

2. Explain why 1.21 cannot be the probability of some event.

Answer

The value 1.21 is larger than 1.00. Again, any valid probability value will be between $0.00 = 0\%$ and $1.00 = 100\%$, inclusively.

3. Explain why -0.22 cannot be the probability of some event.

Answer

The value -0.22 is negative. Any valid probability value will be non-negative since the value must be between $0.00 = 0\%$ and $1.00 = 100\%$.

4. Give an example of a situation and an impossible event for that situation.

Answer

Answers to this can vary greatly. An example tied to the situation about rolling pairs of dice would be the event of rolling two standard game dice with a sum of 13 up. Since the largest each die can be is 6, the sum cannot exceed 12; the event is an impossible event. That is, $P(\text{rolling a sum of } 13) = 0.00\%$.

There is one final fundamental property of probability for the events and sample space within any given situation. The **sum of the probabilities** of all possible outcomes within the sample space of a given situation must always total $1.00 = 100\%$.

Basic Methods for Computing Simple Probabilities

With our basic terminology established, we focus on how to compute a given situation's probability. We initially examine three basic and commonly used methods. Our first method is called the **subjective** or **intuitive method,** where we produce a numerical estimate of the probability based on personal judgment, past experiences, or even personal opinion. For example, having purchased a few non-winning lottery tickets in the past, yet hearing of a winner on the news, we might estimate the probability of winning as a low value of 1%. We do not depend on subjectively determined probability values in quality statistical work. Instead, we form more robust and accurate methods for determining these values.

This leads us to our second method, commonly called the **classical method.** In this method, if a given situation has n different equally likely outcomes in the sample space and if event A can occur in x ways, then

$$P(A) = \frac{\text{number of ways event } A \text{ can occur}}{\text{number of outcomes in the sample space}} = \frac{x}{n}.$$

As an example, if a standard die is fair (so each face of the die has equal chance of landing up when the die is rolled), then

$$P(\text{ROLLING A THREE}) = \frac{\text{number of ways rolling a three can occur}}{\text{number of outcomes in the sample space of rolling a die}} = \frac{1}{6} \approx 0.1667 = 16.67\%$$

Notice that this outcome would not be considered unusual since the probability measure is over 5%.

? Text Exercise 3.1.3

Consider the situation in which a pair of fair dice are rolled. Find the probability of each event given. Write results in probability notation and determine if the outcome is considered unusual.

1. ROLLING A THREE ON THE FIRST DIE AND A FOUR ON THE SECOND DIE.

Answer

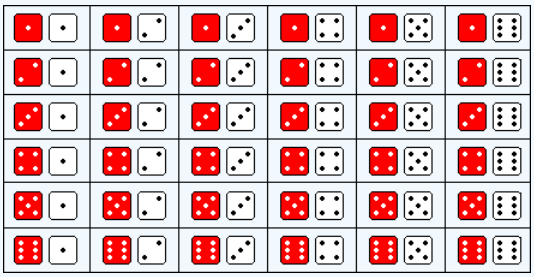


Figure 3.1.3 Sample space of rolling a pair of standard game dice

Since the dice are said to be fair dice, each of the thirty-six outcomes shown in our sample space above is equally likely. We will use the classical approach to determining the probability, using A to represent our event of ROLLING A THREE ON THE FIRST DIE AND A FOUR ON THE SECOND DIE. We notice only one outcome in the sample space matches the event description. Therefore, $P(A) = \frac{1}{36} \approx 0.0278 = 2.78\%$. Since this probability measure is less than 5%, then for our course, we will consider this outcome as unusual.

2. ROLLING A THREE ON ONE DIE AND A FOUR ON THE OTHER DIE.

Answer

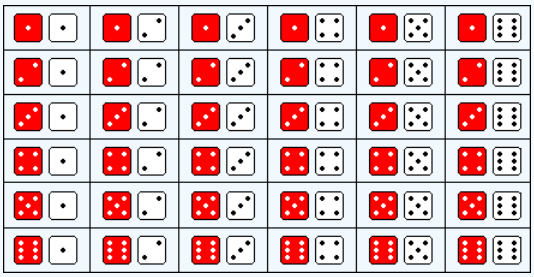


Figure 3.1.4 Sample space of rolling a pair of standard game dice

Using similar reasoning as above and using B to represent our event of ROLLING A THREE ON ONE DIE AND A FOUR ON THE OTHER DIE, notice that there are two outcomes in the sample space that match the event description. Therefore, $P(B) = \frac{2}{36} = \frac{1}{18} \approx 0.0556 = 5.56\%$. Since this probability measure is more than 5%, then for our course, we will not consider this outcome as unusual.

3. ROLLING TWO DICE IN WHICH THE SUM OF THE NUMBER IS SIX.

Answer

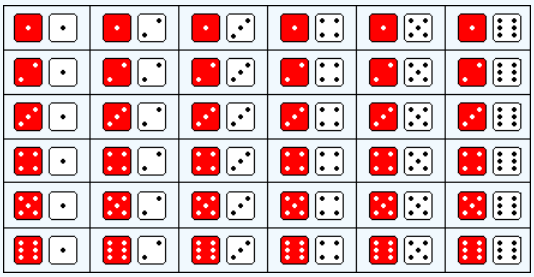


Figure 3.1.5 Sample space of rolling a pair of standard game dice

Let C represent our event of ROLLING TWO DICE IN WHICH THE SUM OF THE NUMBER IS SIX, we notice that there are five outcomes in the sample space that match the event description...can you find all of these? Therefore, $P(C) = \frac{5}{36} \approx 0.1389 = 13.89\%$. Since this probability measure is more than 5%, then for our course, we will not consider this outcome as unusual.

Our third method of computing probabilities is the **empirical, experimental,** or **relative frequency method**. In this method, we repeatedly conduct an experiment, noting the outcomes of the trials, to establish an estimate of probability measures. In repeating a given experiment n times, and noting event A occurred f times, then

$$P(A) = \frac{\text{number of times event } A \text{ occurred}}{\text{number of times situation/experiment was repeated}} = \frac{f}{n}.$$

Notice how this method relates to our previous work producing relative frequency distributions when summarizing data sets. In a sense, we were producing probability measures with our relative frequency values.

The quality of the probability estimate is dependent on the number of repeated trials used. For example, a researcher finds that 25 of 150 randomly selected Kansas teens texted while driving during the last week, empirically indicating that the proportion of Kansas teens that drove while texting last week is $\frac{25}{150} \approx 16.67\%$. Equivalently, there is approximately a 16.67% probability that a randomly selected Kansas teen drove while texting last week. This measure based on only 150 teens does not give us high confidence in this estimated probability value; if the researcher could collect data from $1,500$ Kansas teens, we could be more confident in the estimation.

In general, we require the number of carefully designed repeated trials to be as large as possible to produce an estimate of a probability value. Two underlying assumptions must be used with the experimental frequency approach. First, if an event occurred with a certain probability in past trials, this same event will occur about the same percentage of times in future trials. Second, the relative frequency probability of an event will tend to approach the true probability value as more and more trials are measured (this is commonly referred to as the **Law of Large Numbers**).

There are times when simulations (especially computer simulations) produce a large number of trials and a reasonably accurate measure of the true probability of specific events, especially when the outcomes in a situation are not equally likely. Much weather forecasting is based on computer simulation of outcomes based on regional weather conditions. Similarly, the spread of infectious diseases is modeled by computer simulations to predict outcomes while avoiding extensive medical research costs or without impacting living organisms. For example, suppose we wonder if our game die is fair--that each face has an equal likelihood of occurring when the die is rolled. We begin a simulation with the assumption that the die is fair, but upon rolling the die five times, we roll a one three of the five times--indicating an empirical probability of $P(\text{ROLL OF ONE}) = \frac{3}{5} = 60\%$. Of course, only rolling five times is insufficient for us to have high confidence in the accuracy of the empirical probability measure. We continue rolling the die three hundred times, in which we roll a one 188 times. This new empirical probability measure of $P(\text{ROLL OF ONE}) = \frac{188}{300} \approx 62.67\%$ indicates that our die is not likely to be fair; the experimental probability for the one is not the expected fair value of $P(\text{ROLL OF ONE}) = \frac{1}{6} \approx 16.67\%$.

? Text Exercise 3.1.4

In 1856, Gregor Mendel began to study different inherited features, such as the color of pea plants. According to one source, in a second-generation group of pea plants, 6002 peas produced by the plants were yellow, and 2001 were green in color. What was the empirical probability that a randomly selected second-generation pea would be green? Is this close to the hypothesized value that Mendel claimed of 25%?

Answer

We compute that $P(\text{GREEN COLORED PEA}) = \frac{2001}{6002+2001} \approx 25.0031\%$.Mendel's experimental probability is extremely close to his hypothesized probability claim of 25%.

Complement and Complement Probabilities

There are times we will be interested in finding the probability that an event A does not occur. The collection of all outcomes in a sample space in which a given event A does not occur is called the **complement** of event A and is denoted by $\bar{A}$. Other sources may use different notations to denote the complement; common ones include A^c or $\sim A$. The idea of complementary events allows us to divide the sample space into two mutually exclusive groups (no outcome can be found in both of the two groups A and $\bar{A}$) and also exhaustive (every outcome of the sample space must be included in one of our two groups). For example, suppose A is the event of getting two different numbers on each die when two dice are rolled; there are 30 such outcomes in our sample space that

meet this event description. $\bar{A}$ is the event of getting the same numbers on each dice when two dice are rolled; there are 6 such outcomes in our sample space. Between A and $\bar{A}$, all 36 outcomes appear exactly once.

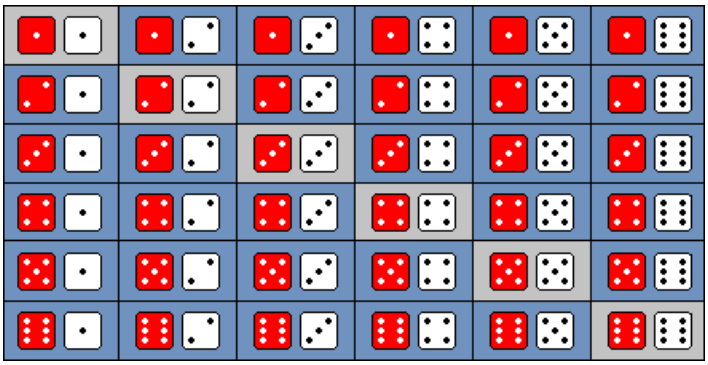


Figure 3.1.2: Events A (blue background) and $\bar{A}$ (grey background)

Since all outcomes of a situation must be either in the event or the complement of the event, we have the following three key consequences:

1. $P(A)+P\left(\bar{A}\right) = 1 = 100\%$
2. $1-P(A) = P\left(\bar{A}\right)$
3. $1-P\left(\bar{A}\right) = P(A)$

Notice that the first consequence leads naturally to the other two with basic algebra. Suppose we wish to determine the probability of event A from above ($P(\text{GETTING TWO DIFFERENT NUMBERS OF PIPS ON EACH DIE WHEN TWO DICE ARE ROLLED})$). We can go back to our sample space to count all such outcomes or use the complement concept to produce our results quickly. We have

$$\begin{aligned} P(A) &= 1-P\left(\bar{A}\right) \\ &= 1-\frac{6}{36} \\ &= \frac{30}{36} = \frac{5}{6} \approx 0.8333 \text{ or } 83.33\% \end{aligned}$$

This example illustrates that it is sometimes easier to determine the probability of a complement event instead of the given event. Once we know the probability of the complement event, we can easily determine the probability of the event.

? Text Exercise 3.1.5

Use the concepts of complement events to answer the questions below:

1. Suppose there are sixty tiles in a bag of which 13 are green, 6 are yellow, 4 are pink, 9 are red, 8 are purple, and 20 are black. The tiles are well-mixed. We will randomly draw one tile from the bag without looking into the bag. We want to determine the probability that we draw a tile that is not a primary color; that is, we are to find $P(\text{NOT A PRIMARY COLORED TILE})$.

Answer

As a reminder, there are three primary colors: red, yellow, and blue. Although the complement is not necessary to answer this question, the use of the complement makes the problem easier to compute. Since "a primary colored tile" is the complement event of "not a primary colored tile" in this situation, we notice that $P(\text{NOT A PRIMARY COLORED TILE}) = 1-P(\text{A PRIMARY COLORED TILE}) = 1-\frac{6+9}{60} = \frac{45}{60} = 75\%$.

2. A number is chosen randomly from the set of integers between 1 and 99, inclusively. What is the probability of randomly selecting a number that is not a perfect square?

Answer

We notice that there are only a few integers inclusively between 1 and 99 that are perfect squares and many that are not. Specifically the perfect square integers in this set are $\{1, 4, 9, 16, 25, 36, 49, 64, 81\}$. Thus $P(\text{NOT A PERFECT SQUARE}) = 1 - P(\text{PERFECT SQUARE}) = 1 - \frac{9}{99} \approx 0.909 = 90.9\%$.

3. Suppose we want to know the probability of randomly selecting a group of 35 people in which at least two people will have the same birth day in a year (such as September 1^{st} or May 28^{th} -- we will ignore leap years for simplicity). Suppose we also know the probability of randomly selecting a group of 35 people, in which no two people have the same birth day in the year, which is about 18.56%. (We will discuss how this value of 18.56% can be found later in the chapter). From this information, can we determine the probability of randomly selecting a group of 35 people in which at least two will have the same birth day in a year?

Answer

We see that the complement of "at least two people will have the same birth day in a year" is the description "less than two people (that is none) will have the same birth day in a year." We can use our complement rule to note that $P(\text{AT LEAST TWO WILL HAVE THE SAME BIRTHDAY}) = 1 - P(\text{NO TWO WILL HAVE THE SAME BIRTHDAY}) = 1 - 0.1856 = 0.8144 = 81.44.\%$

There are times when clearly describing the complement of an event can be simple. For example, in the roll of a single standard game die, it is common to quickly describe the complement of AN EVEN NUMBER ON A ROLL to be AN ODD NUMBER ON A ROLL. There are also times when clearly describing the complement of an event can be challenging. We must ensure the complement description covers all possible outcomes for a situation. For a different example, in random selection of a number from the real number line, it is common to quickly describe the complement of SELECTION OF A NEGATIVE NUMBER as SELECTION OF A POSITIVE NUMBER. However, there is the number, 0, that is neither positive nor negative which means we have an incorrect complement description. The true complement to SELECTION OF A NEGATIVE NUMBER is SELECTION OF A NON-NEGATIVE NUMBER. Now all possible outcomes have been accounted for in the descriptions.

We must think carefully about our sample space and correctly identify all outcomes which are not in our event. As another example, we might have the situation of randomly selecting an FHSU student with a focus on the outcome of STUDENT HAS HAD AT LEAST THREE COVID VACCINE SHOTS. The complement is the outcome of STUDENT HAS HAD TWO OR FEWER COVID VACCINE SHOTS or, equivalently, STUDENT HAS HAD FEWER THAN THREE COVID VACCINE SHOTS. Notice how the use of mathematical notation can help us here. We might represent the outcome of STUDENT HAS HAD AT LEAST THREE COVID VACCINE SHOTS more briefly in notation as $x \geq 3$. Then, in the context of counting number of shots, the complement is $x < 3$, which leads to a proper complement description of STUDENT HAS HAD FEWER THAN THREE COVID VACCINE SHOTS.

This leads to special cases that commonly cause problems in complement descriptions; specifically, events involving "at least," "at most," "all' or "none." The complement of "all are" is not "none are," but is instead "at least one is not." For example, the complement of STUDENT HAS HAD ALL AVAILABLE COVID VACCINE SHOTS would be STUDENT HAS MISSED AT LEAST ONE COVID VACCINE SHOT. The complement of STUDENT HAS HAD NONE OF THE AVAILABLE COVID VACCINE SHOTS would be STUDENT HAS HAD AT LEAST ONE OF THE AVAILABLE COVID VACCINE SHOTS. Note that the complement of "none are" is "at least one is." We must think carefully when dealing with complements of event claims involving such keywords.

? Text Exercise 3.1.6

Give written descriptions of the complements of each event described below

1. The event: IT SNOWS ON CHRISTMAS DAY

Answer

The complement event description would be IT DOES NOT SNOW ON CHRISTMAS DAY. We note that IT RAINS ON CHRISTMAS DAY is an outcome that fits in the complement description, but is itself not the actual full complement since other outcomes exist in the complement besides rain.

2. The event: NATASHA IS LESS THAN 10 MINUTES LATE

Answer

Notice the given event description of less than 10 minutes can be represented mathematically as $x < 10$, informing us that the complement must be related to $x \geq 10$. Therefore, the complement event description would be NATASHA IS AT LEAST 10 MINUTES LATE, or equivalently, NATASHA IS LATE BY AT LEAST 10 MINUTES. The phrasing we might give can vary, but the meaning of the phrasing must be tied to the inequality $x \geq 10$.

3. The event: ALL CARDS IN A POKER HAND ARE FACE CARDS (Click here for a full description and visualization of a standard deck of playing cards)

Answer

The given event description of "all cards" is complemented by "at least one is not." So the complement event description would be AT LEAST ONE CARD IS NOT A FACE CARD. We must think clearly about this complement and also note why the description NONE OF THE CARDS ARE FACE CARDS is not the complement description. We could have one, two, three, or four cards that are not face cards as possible events in the complement.

4. The event: NONE OF THE STUDENTS FAILED THE LAST EXAM

Answer

The given event description of "none failed" is complemented by "at least one did fail." So the complement event description would be AT LEAST ONE STUDENT FAILED THE LAST EXAM.

Summary

To review this section, we list several important facts to remember when working with probabilities:

1. An outcome of a random experiment is any potential result of the experiment. An event is a set of outcomes one might expect from a given random experiment. The sample space is the collection of all possible outcomes in a given situation.
2. The probability of an event A is denoted by $P(A)$ with the condition that $0 \leq P(A) \leq 1$. If $P(A) \leq 5\%$, we will currently consider the event as unusual.
3. The sum of the probabilities for all the possible outcomes in the sample space will always total to $1 = 100\%$.
4. Multiple methods may be used for computing probability values. We have discussed the methods of "subjective/intuition," "classical for equally likely simple events," and "empirical/experimental/relative frequency." Some methods only produce estimations of the actual probability value.
5. The complement of an event A is denoted by $\bar{A}$ and must contain all outcomes of the sample space that are not part of the given event. As such, we have the probability benefit of $P(A) + P(\bar{A}) = 1$ that can be used to find the probability of an event if we know the probability of the complement.

3.2: Counting Strategies

Learning Objectives

- Use tree diagrams to organize outcomes in a series of activities
- Develop and use the Multiplication Rule for counting the number of possible outcomes, including the factorial method
- Develop and use the permutation method for counting the number of possible outcomes in ordered arrangements
- Develop and use the combination method for counting the number of possible outcomes in unordered arrangements

Outcomes from a Series of Activities

We have discussed the classical method of computing probabilities for an event A by the formula

$$P(A) = \frac{\text{number of ways event } A \text{ can occur}}{\text{number of outcomes in the sample space}} = \frac{x}{n},$$

provided outcomes in the sample space are equally likely. The formula requires determining the total number of outcomes in the sample space and in A. Up to this point, our examples and exercises have dealt with reasonably small sample spaces; however, the number of outcomes of interest may be quite large. This section develops counting techniques that help us in determine the number of possible outcomes. We begin by finding a logical way to organize the development of a sample space for multi-trial situations, such as tossing multiple dice. Ultimately, we will develop several counting methods.

Tree Diagrams and the Multiplication Rule for Counting

A **tree diagram** helps us list the various outcomes in a series of activities. To build a tree diagram, we list all outcomes of the first activity. Next to each of those listed outcomes, we branch to list all outcomes of the second activity. We branch off each of the second listed outcomes for the third activity, and so on. Eventually, all activities in the situation will be completed. Let us use a familiar example to build our first tree diagram. Suppose we wish to develop the sample space for rolling two dice. First, we list all the outcomes that may occur with the first die roll (the first activity of the described situation).

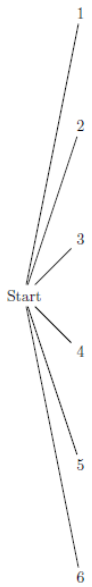


Figure 3.2.1: Tree diagram of first die roll

Next, from each of those six listed outcomes, we list all the outcomes that may occur with the second die roll (the second activity of the described situation).

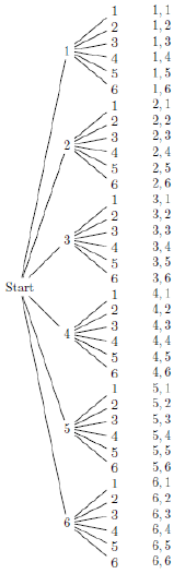


Figure 3.2.2: Tree diagram of rolling two dice

Notice that, at the right of the diagram, we listed all thirty-six branches of the tree. This gives us an organized listing of the entire sample space. This branching idea hints at a counting strategy to determine the number of objects in the sample space. There are six branches for the first die roll and six branches from each for the second die roll; this leads to a total of $6 \cdot 6 = 36$ outcomes in the sample space. This strategy is called the **Multiplication Rule for Counting**: if there are m possible outcomes for a first activity A_1 and there are n possible outcomes for a second activity A_2, there are a total of $m \cdot n$ possible outcomes for the experiment in which activity A_1 is followed by activity A_2. This rule extends for any finite number of steps in a situation.

? Text Exercise 3.2.1

Create tree diagrams to determine the entire sample space of the following situations. Be sure to list the final sample space and then note the number of outcomes in the sample space.

1. A food truck allows customers to create a meal by selecting one item from each category: burger, side, and drink. There are three burgers to choose from (B_1, B_2, B_3), two sides (S_1, S_2), and three drinks (D_1, D_2, D_3). Create the sample space of different meals that can be ordered from this food truck.

Answer

We create the tree diagram as shown below. Notice that our first level of branches are the three burger choices, the second level of branches are the two side choices, and the third level of branches are the three drink choices:

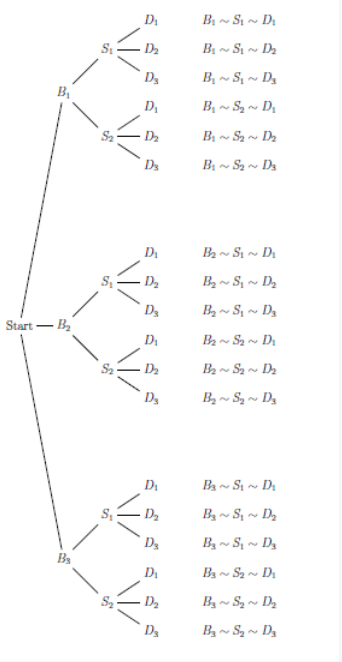


Figure 3.2.3 Tree diagram of food truck meals

Notice that our diagram gives us our sample space.

$$\left\{\begin{array}{c} B_1 \sim S_1 \sim D_1,\ B_1 \sim S_1 \sim D_2,\ B_1 \sim S_1 \sim D_3, \\ B_1 \sim S_2 \sim D_1, B_1 \sim S_2 \sim D_2,\ B_1 \sim S_2 \sim D_3, \\ B_2 \sim S_1 \sim D_1,\ B_2 \sim S_1 \sim D_2,\ B_2 \sim S_1 \sim D_3, \\ B_2 \sim S_2 \sim D_1, B_2 \sim S_2 \sim D_2,\ B_2 \sim S_2 \sim D_3, \\ B_3 \sim S_1 \sim D_1,\ B_3 \sim S_1 \sim D_2,\ B_3 \sim S_1 \sim D_3, \\ B_3 \sim S_2 \sim D_1, B_3 \sim S_2 \sim D_2,\ B_3 \sim S_2 \sim D_3 \end{array}\right\}$$

The sample space consists of 18 total outcomes. The Multiplication Rule of Counting predicts the total number of outcomes to be $3 \cdot 2 \cdot 3 = 18$.

2. There are six tiles well-mixed in a bag. The tiles are identical except in color; three are red, two are white, and one is blue. We are to randomly draw two tiles from the bag, one at a time, without replacement. Produce the sample space of two tiles that can be drawn from the bag.

Answer

Again, we can create the diagram, but we must think clearly about each step of tile selection since the choices available for the second tile selection depend on what happened in the first tile selection. Since the tiles are identical, we cannot distinguish between those tiles of the same color. We might draw any of the three colors in the first tile selection, but in the second tile selection, we cannot select blue if we have selected blue on the first draw. We can organize our listing to produce the following tree diagram.

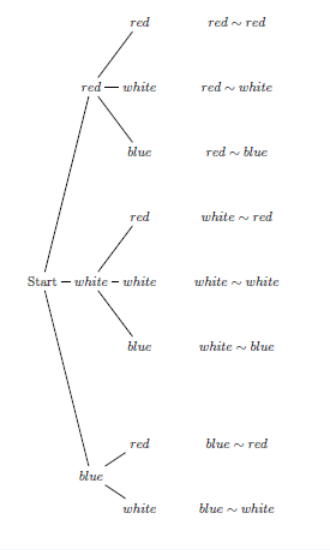


Figure 3.2.4 Tree diagram of tile combinations

The final sample space is

$$\left\{\begin{array}{c} red \sim red,\ red \sim white, red \sim blue, \\ white \sim red,\ white \sim white, white \sim blue, \\ blue \sim red,\ blue \sim white \end{array}\right\}$$

which includes 8 total outcomes. Notice how our Multiplication Rule of Counting must be carefully used because we cannot select the blue tile again. There were two colors (specifically red and white) that, if selected first, still allowed three colors to be chosen in the second selection. There was one color (specifically blue) that, if selected first, only allowed two colors to be chosen in the second selection. We can adjust the rule's application so that $2\dot{3}+1 \cdot 2 = 8$ predicts the total number of outcomes in the sample space.

3. A salesman must visit four important clients: Mr. White, Miss Scarlet, Professor Plum, and Mrs. Peacock. Determine the number of choices the salesman has to visit these clients.

Answer

We create the tree diagram as shown below, which gets a bit large. (One thing we might consider is to give codes to the four clients, such as numbering them 1 to 4 and using the numbers instead of their longer names.) The tree is reasonably easy to produce if we maintain our organization while creating each branch level, one at a time.

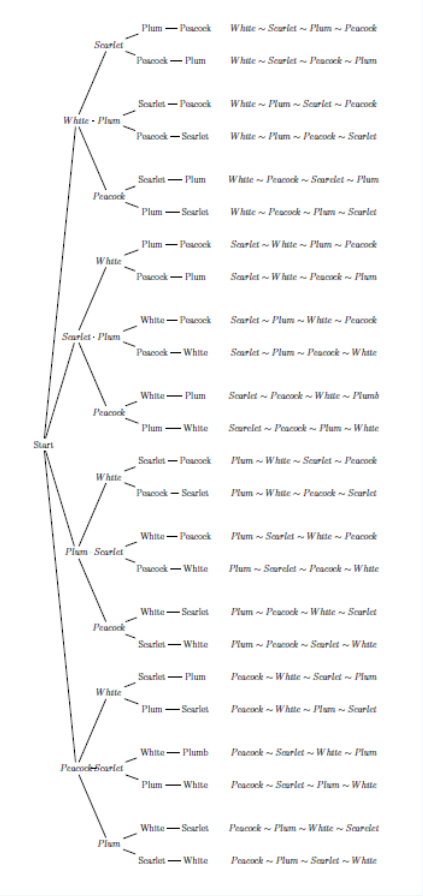


Figure 3.2.5 Tree diagram of client visits

In this situation, we will not list the sample space of 24 distinct outcomes since it can be seen in the last column of the tree diagram. Notice how our Multiplication Rule of Counting predicts the number of outcomes at $4 \cdot 3 \cdot 2 \cdot 1 = 24$. We have 4 options for the first visit, 3 options for the second, 2 options for the third, and then only 1 options for the final visit.

In our examples of tree diagrams, we can begin counting the number of outcomes by using our multiplication rule, but we must consider what happens at each branching action of the tree. Once we become comfortable with the concept of the multiplication rule, we can more quickly compute the number of possible outcomes.

First Basic Counting Techniques

We have shown that the Multiplication Rule for Counting is a valuable tool for determining the number of possible outcomes in many situations. We now use this tool to develop general counting techniques without producing tree diagrams. As we develop various techniques, remember that all the methods are developed from the Multiplication Rule for Counting.

Suppose June buys a new cell phone and must choose a six-digit code to keep it secure when she is not using it. How safe is her phone from someone who randomly chooses six digits to unlock it? Knowing that June will select one, we need to determine the number of possible six-digit codes. The standard convention in choosing a code allows for digits to repeat. Later, we will see what happens if the code cannot use digits more than once. We can think through the branching of a tree diagram using a simple counting diagram to determine the number of outcomes possible,

# of Possibilities	___	___	___	___	___	___	
Choices for	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	Code digit position

where the row of blanks represent the numbers to be chosen in the given position in the code. We only need to place the number of possibilities in each blank for each code position. Recall that there are ten digits, 0 to 9, and with use of the multiplication rule, we have

$$\begin{array}{lcccccc} \#\text{ of Possibilities} & \underline{10} & \underline{10} & \underline{10} & \underline{10} & \underline{10} & \underline{10} \\ \text{Position} & 1^{st} & 2^{nd} & 3^{rd} & 4^{th} & 5^{th} & 6^{th} \end{array} \quad \Rightarrow 10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 10^6 = 1,000,000$$

There are $1,000,000$ unique codes that June could choose from. The probability of someone randomly guessing June's code is $P(\text{GUESSING CODE}) = \frac{1}{1,000,000} = 0.000001$. This event is highly unlikely; June should feel relatively safe about someone randomly guessing her code. We would not want to construct the sample space in this situation due to the number of outcomes possible, yet we can still reflect on the sample space and know how many codes are in the sample space.

In Text Exercise $3.2.1.3$ regarding a salesman visiting four distinct clients, we can use a similar approach:

$$\begin{array}{lcccc} \#\text{ of Possibilities} & \_\_\_ & \_\_\_ & \_\_\_ & \_\_\_ \\ \text{Choices for} & 1^{st} & 2^{nd} & 3^{rd} & 4^{th} \end{array} \quad \text{Client visit position}$$

where the row of blanks represent the order to visit the four clients. We know that we have four clients to choose from in the first position; once that choice is made, there are three clients left to choose from for the second position; similarly, there are just two clients left to choose for the third position; and finally, only one client will be left for the fourth position:

$$\begin{array}{lcccc} \#\text{ of Possibilities} & \underline{4} & \underline{3} & \underline{2} & \underline{1} \\ \text{Position} & 1^{st} & 2^{nd} & 3^{rd} & 4^{th} \end{array} \quad \Rightarrow 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

There are 24 unique orderings the salesman could choose for visiting the four clients. This last example is a multiplication pattern that frequently occurs in counting outcomes. The shortened notation for this computation is called the **factorial notation** $4!$. The $!$ symbol is read "factorial", $4!$ is read "four factorial", and $4!$ means, computationally, $4 \cdot 3 \cdot 2 \cdot 1$. In general $n!$ represents the product $n(n-1)(n-2) \cdot \ldots \cdot 2 \cdot 1$. We assign $0! = 1$ by special definition. Notice how quickly factorials increase in value; for example $8! = 40,320$ and $20! = 2.4329 \times 10^{18}$. Try the text exercises below for more practice with counting the number of possible outcomes.

? Text Exercise 3.2.2

Determine the total number of outcomes possible for each of the given descriptions.

1. We have five standard fair coins of different values: a penny, a nickel, a dime, a quarter, and a half-dollar. When tossed, each lands on heads (H) or tails (T). One way they might land, in order of increasing value, is HHTHT.
   a. How many ways can the coins land when tossed and placed in order of value, say from smallest value to largest?
   b. How many ways can we order the five different coins?

Answer

Both parts of this question ask for a count, and we use the multiplication rule for each, though the results are different.

a. We establish our simple counting diagram, remembering that the coins are placed in order of increasing value:

$$\begin{array}{lccccc} \#\text{ of Possibilities} & \_\_\_ & \_\_\_ & \_\_\_ & \_\_\_ & \_\_\_ \\ \text{Position of Coins} & 1^{st} & 2^{nd} & 3^{rd} & 4^{th} & 5^{th} \end{array} \quad \begin{array}{c} \text{each coin's landing side} \\ \text{by denomination} \end{array}$$

then we notice that each trial has two possible outcomes, either H or T. Our counting diagram becomes

$$\begin{array}{lccccc} \#\text{ of Possibilities} & \underline{2} & \underline{2} & \underline{2} & \underline{2} & \underline{2} \\ \text{Position} & 1^{st} & 2^{nd} & 3^{rd} & 4^{th} & 5^{th} \end{array} \quad \Rightarrow 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^5 = 32$$

This tells us there are 32 possible head and tail arrangements.

b. Next, we examine the number of ways the five distinct coins can be placed in an order, such as from smallest value to largest, largest to smallest, or even some other arrangement. It is important to note that each of the coins is distinguishable from the other, so there is no confusion on which coin is in which location. We establish our simple counting diagram as an aide:

$$\begin{array}{lccccc} \#\text{ of Possibilities} & \_\_\_ & \_\_\_ & \_\_\_ & \_\_\_ & \_\_\_ \\ \text{Position in Arrangement} & 1^{st} & 2^{nd} & 3^{rd} & 4^{th} & 5^{th} \end{array} \quad \text{by coin denomination}$$

and see there are five choices for a coin to be placed in the 1^{st} position. Once that coin is chosen, we have only four choices for the coin in the second position, and so on. Our counting diagram becomes

$$\begin{array}{ll} \#\text{ of Possibilities} & \underline{5} \quad \underline{4} \quad \underline{3} \quad \underline{2} \quad \underline{1} \\ \text{Position} & 1^{st} \quad 2^{nd} \quad 3^{rd} \quad 4^{th} \quad 5^{th} \end{array} \quad \Rightarrow 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5! = 120$$

This tells us there are 120 possible arrangements of the coins by denomination. In the upcoming optional section, we will revisit this situation, in which some of the coins are not distinguishable from each other, requiring adjustments in our counting method.

2. There are two distinctly colored standard decks of playing cards (red and blue). How many different outcomes are possible if we draw four cards in sequence, with two cards drawn first from the red deck and then two more from the blue deck?

Answer

We again establish our simple counting diagram:

$$\begin{array}{lllll} \#\text{ of Possibilities} & \_\_\_ & \_\_\_ & \_\_\_ & \_\_\_ \\ \text{Choices for} & 1^{st} & 2^{nd} & 3^{rd} & 4^{th} \\ & \text{red} & \text{cards} & \text{blue} & \text{cards} \end{array} \quad \text{each card's position}$$

then we determine the possible outcomes for each card position. Our counting diagram becomes

$$\begin{array}{ll} \#\text{ of Possibilities} & \underline{52} \quad \underline{51} \quad \underline{52} \quad \underline{51} \\ \text{Position} & 1^{st} \quad 2^{nd} \quad 3^{rd} \quad 4^{th} \\ & \text{red} \quad \text{cards} \quad \text{blue} \quad \text{cards} \end{array} \quad \Rightarrow 52 \cdot 51 \cdot 52 \cdot 51 = 7,033,104$$

That is over 7 million different possible outcomes! We are definitely glad we did not have to produce the entire sample space for this simple little situation.

3. Leisha is a research biologist for a sod (grass) company. She must research the effects of four fertilizer types on three temperature zones with three watering amounts. How many different sod plots must she create to test all possible configurations of these three factors?

Answer

We jump to the final computing action to tell Leisha how many plots she will need to perform her research:

$$\begin{array}{ll} \#\text{ of Possibilities} & \underline{4} \quad \underline{3} \quad \underline{3} \\ \text{Factor Choice} & \text{fertilizer} \quad \text{temp. zones} \quad \text{water amts.} \end{array} \quad \Rightarrow 4 \cdot 3 \cdot 3 = 36$$

Leisha will need to have 36 different plots for her desired sod research.

4. Carl is a psychologist studying telepathy between patients who claim such abilities. He will test their claims using a deck of six cards, each with a different picture. The cards will be shuffled randomly before each pair of patients uses the cards in his experiment. How many different card arrangements (shuffled decks) are possible?

Answer

Again, we jump to the final computing action to answer Carl's dilemma:

$$\begin{array}{ll} \#\text{ of Possibilities} & \underline{6} \quad \underline{5} \quad \underline{4} \quad \underline{3} \quad \underline{2} \quad \underline{1} \\ \text{Card deck position} & 1^{st} \quad 2^{nd} \quad 3^{rd} \quad 4^{th} \quad 5^{th} \quad 6^{th} \end{array} \quad \Rightarrow 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 6! = 720$$

So Carl will have $6! = 720$ different card arrangements possible in the deck of six cards. We mention that this is a factorial count situation. Once we recognize a factorial count, we will reach the final answer of $6!$ with no diagram work. We note for Carl that if the probability of one patient in the study randomly guessing all six cards correctly as examined only by their telepathy partner is $\frac{1}{720} \approx 0.1389\%$-not impossible but highly improbable.

Counting with Permutations

Suppose, in the last problem from Text Exercise 3.2.2, Carl selected only three of the six cards for his patients to test their telepathic abilities. We can still use the same method described above.

$$\begin{array}{lccc} \#\text{ of Possibilities} & \underline{6} & \underline{5} & \underline{4} \quad \Rightarrow 6\cdot 5\cdot 4 = 120 \\ \text{Card deck position} & 1^{st} & 2^{nd} & 3^{rd} \end{array}$$

Carl will have 120 different smaller deck arrangements if only selecting any three of the six original cards for the smaller deck.

More formally, we call situations that are counted this way permutations. A **permutation** count is the number of ways to arrange, in order, n distinct objects, taking r at a time. We denote this symbolically as ${}_nP_r$ where n and r are whole numbers with $n \geq r$. The two general forms for computing this count are given by:

$$\begin{aligned} {}_nP_r &= n(n-1)(n-2)\cdots(n-r+1) && (1) \\ &= \frac{n!}{(n-r)!} && (2) \end{aligned}$$

When building the counting action, we tend to use (1). When working with permutations theoretically, we tend to represent permutation counts using (2). We should become familiar with both forms of computing permutations. Using Carl's smaller deck we can see the computation as

$${}_6P_3 = n(n-1)(n-2)\cdots(n-r+1) = 6\cdots(6-3+1) = 6\cdot 5\cdot 4 = 120$$

and also

$${}_6P_3 = \frac{n!}{(n-r)!} = \frac{6!}{(6-3)!} = \frac{6\cdot 5\cdot 4\cdot \cancel{3}\cdot \cancel{2}\cdot \cancel{1}}{\cancel{3}\cdot \cancel{2}\cdot \cancel{1}} = 120$$

Our fraction form of computation with factorials is canceling out the unused $(n-r)!$ objects in the count from the $n!$. We also want to quickly recognize situations requiring permutation counts (especially when large numbers are involved) so we can use technology to perform the calculations.

? Text Exercise 3.2.3

Determine the number of outcomes possible in the described situations:

1. Matching questions are sometimes used in statistics exams. If ten distinct statistical symbols are to be matched uniquely with ten distinct statistical terms, how many symbols-to-terms matches (correct or incorrect) are possible?

Answer

We could use a counting diagram, but we produce the counts by recognizing the strategies discussed without using a diagram in these answers. Thinking about our statistical term choices for each of the ten symbols, we see there are $10\cdot 9\cdot 8\cdots 2\cdot 1 = 10! = 3,628,800$ different symbol-to-term matches students can make. A student strictly randomly guessing all ten correctly has a probability of only $\frac{1}{3,628,800} \approx 0.0000002756$ We recognize that this count is not only a simple factorial count but that any factorial count can be considered a permutation count. Specifically $10! = {}_{10}P_{10} = \frac{10!}{(10-10)!}$.

2. In the same matching question used above, if there are only seven distinct statistical symbols to be matched with the given ten distinct statistical terms, how many symbols-to-terms matches might be made by students?

Answer

In this case, we only match seven symbols (places) with ten terms (objects). There are $10\cdot 9\cdots 4 = {}_{10}P_7 = 604,800$ different symbol-to-term matches that students can make. After a little thought, we should recognize that this is a permutation counting situation since the order is important.

3. A saleswoman must visit only four of her nine clients tomorrow. How many different ordered client visit lists can the saleswoman make?

Answer

At this point, we recognize quickly that this is also a permutation count since order matters. The saleswoman has ${}_9P_4 = 3,024$ ordered client visit lists that she might make. We doubt she will make all these, but knowing how many are possible speaks to the variety of ordered lists she has available. We can also use technology to compute such permutation values after we note the values of n and r. For example, in an Excel spreadsheet cell, we can enter $=\text{PERMUT}(9,4)$ and the value 3024will be shown.

4. During a two-hour time frame, a nurse must take vital signs measurements from ten of his forty assigned patients. How many different lists can the nurse make for an ordered round involving ten of the forty patients?

Answer

We quickly recognize that this is also a permutation count. The nurse has ${}_{40}P_{10}$ ordered patient lists that he might make for his two-hour round. We will use technology for the computation: in an Excel spreadsheet cell, we can enter $=\text{PERMUT}(40,10)$ producing the approximate value $3.07599\text{E+}15$; this is a standard technology representation of the number 3.07599×10^{15}.He can make well over a quadrillion ordered patient lists in this situation. We might be amazed at the possibilities and appreciate how many randomly chosen ordered lists can be made.

5. We return to our "Birthday" problem of Text Exercise $3.1.5$,which asks, "What is the probability that in a random group of thirty-five people, two or more of the group will share a birth day in the year (such as September 1^{st} or May 28^{th})?" To answer this question, we assume 365 days in a year (ignore leap year cases), that each day of the year is equally likely for births, and the birth day of the various individuals in the group is not dependent on another individual in the group.

Answer

We first note that in a group of thirty-five people placed in some order, the first person will have one of 365 different birth days, the second will have one of 365 birth days, ... ,the thirty-fifth will have one of 365 birth days. There are $365^{35} \approx 4.7891 \times 10^{89}$ possible lists of thirty-five birth days that various groups of thirty-five people can form.

Now, we must find how many of those lists have two or more with the same birth day in the year. To do so, we must look at many separate cases: how many of our lists have exactly two in a birth day list the same or how many have precisely three the same or ... or how many have all thirty-five with the same birth day or how many have two different pairs the same or...etc. But being reflective on the work in finding so many separate counts and being wise to consider other methods, we again notice that there is only one case in the complement description: having lists in which none of the thirty-five have the same birth date. The use of the complement method makes this problem approachable. Therefore, we must count only the number of lists of birth days that can be made in which none of the thirty-five are the same birth day.

Since this is an ordered list, we recognize this is a permutation count in which we are forming lists of 35 distinct birth days coming from 365 options...that is, we need ${}_{365}P_{35} = \frac{365!}{(365-35)!} \approx 8.8893 \times 10^{88}$. So, $P(\text{NONE HAVE THE SAME BIRThDAY}) \approx \frac{8.8893 \times 10^{88}}{4.7891 \times 10^{89}} \approx 0.1856$.

Now, the $P(\text{AT LEAST TWO WITH SAME BIRTHDAY}) = 1 - P(\text{NONE HAVE THE SAME BIRTHDAY})$ $\approx 1 - 0.1856 = 0.8142 = 81.42\%$. Notice that in a randomly selected group of thirty-five people, there is a reasonably high probability that at least two will share the same birth day.

It is important to note that order is essential in permutations. Choosing object a and then object b, in that order, is counted as a different outcome from choosing object b first and then object a. When the order of choice is irrelevant (such as the saleswoman choosing to visit four of her nine clients but not concerned with the order of those of her visits), an adjusted approach must be taken: a counting approach termed as a combination.

Counting with Combinations

In each of our previous counting approaches, the order of selection was important to the counting. However, there are times when order is not important to the outcomes of an experiment, especially when we consider samples taken from a population. We begin with a simple, small example that we will later generalize.

Suppose we have four distinct individuals, Adam (A), Betsy (B), Cathy (C), and Damon (D) and we wish to form all possible sample groups of size three from this population of four. Using our tree approach, we can create the ${}_4P_3 = 24$ ways to arrange these four into ordered groups of three--yes, a permutation count. This list is given by:

ABC	DBC	ACD	ABD
ACB	DCB	ADC	ADB
BCA	BCD	CDA	BDA
BAC	BDC	CAD	BAD
CAB	CDB	DAC	DAB
CBA	CBD	DCA	DBA

The order displayed here is intentionally a bit different from our standard tree approach. Thinking back on our need to count how many sample groups of size three are available, we should note that the various colored collections of the ordered groups produce the same (unordered) sample groups. The first column of six shown in red is different arrangements of the same sample group containing Adam, Betsy, and Cathy; the second column of six in blue contains only one sample group of Betsy, Cathy, and Damon; the third column has the one sample group of Adam, Cathy, and Damon; and the fourth column of Adam, Betsy, and Damon. Given any collection of three individuals, we know there are $3 \cdot 2 \cdot 1 = 6$ ways to arrange those three. There are only $\frac{{}_4P_3}{3!} = \frac{24}{6} = 4$ unique sample groups (unordered groups) of size three from this collection of four individuals.

Is there a predictable pattern to develop a counting formula for the number of distinct unordered groups of size r taken from a collection of n distinct objects? First, we performed a permutation count of ${}_nP_r$ to get the total number of ordered groups of size r taken from the n objects. Then, we had to divide by the number of ways any specific group of size r could be arranged--a factorial value of $r!$--to account for the same collection arranged differently. We computed $\frac{{}_nP_r}{r!}$ to determine the number of distinct combinations of n distinct objects taken r at a time. We call this type of counting a **combination** and denote in symbols as ${}_nC_r$ or as $\binom{n}{r}$; in this text, we will use the ${}_nC_r$ form for consistency. We typically read this as, "n choose r" since it is counting how many ways we can choose r objects from a group of n objects. We can express the formula for combination count in two ways.

$$\binom{n}{r} = {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{(n-r)!r!}$$

We can reason algebraically how the second formula is produced from the first by remembering that ${}_nP_r = \frac{n!}{(n-r)!}$. Many computing technologies also have built-in functions to relieve us from mechanical computation once we have recognized a counting situation as a combination. In the Excel spreadsheet, this function is $= \text{COMBIN}(n, r)$.

We must remember that a combination counts the number of ways to form groups of size r from a collection of size n where order within the groups does not matter. Let us do one more example before a collection of text exercises. Suppose a committee of four students must be formed from a class of thirty-five distinct students. How many committees could be formed? Rearranging the order of four students does not create a different committee. Therefore, we must count using combinations. There are ${}_{35}C_4 = \frac{{}_{35}P_4}{4!} = \frac{35 \cdot 34 \cdot 33 \cdot 32}{4 \cdot 3 \cdot 2 \cdot 1} = 52,360$ different possible committees of size four that could be formed from the thirty-five students. Using the Excel command $= \text{COMBIN}(35, 4)$ produces this same value of $52,360$.

? Text Exercise 3.2.4

Determine the number of outcomes possible in the described situations (warning: not all are combination counts).

1. A construction manager must select eight samples from a group of twenty concrete trucks to test the quality of the concrete mixture. How many different groups of eight can be selected?

Answer

Notice that initially, there are twenty trucks to choose between for the first sample, nineteen for the second, and so on until all eight needed trucks are chosen. However, we also recognize that once a group of eight trucks is selected in a specific order, rearranging those eight trucks in a different order will not create a different sample group. This is a combination counting situation, so we compute 20 choose 8 :

$$_{20}C_8 = \frac{20!}{(20-8)! \cdot 8!} = \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 \cdot 15 \cdot 14 \cdot 13}{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 125,970.$$

There are almost $126,000$sample groups of concrete trucks possible for the construction manager to choose. We could also have computed our value in a spreadsheet using $= \text{COMBIN}(20, 8)$ in a cell, producing this same end value of $125,979$.

2. A dial-faced combination lock produced by a new manufacturer uses three distinct numbers in the lock code between 0 and 39. How many different lock codes are available?

Answer

In this case, we are choosing three numbers from a list of forty$(0$ to $39)$. We also recognize that once a group of three numbers is chosen, rearranging those three numbers will create a different lock code. This is a permutation counting situation, so we compute 40 permute 3:

$$_{40}P_3 = \frac{40!}{(40-3)!} = 40 \cdot 39 \cdot 38 = 59,280.$$

There are $59,280$lock codes for the described lock. We should notice that the term "combination lock" does not use the term "combination" in the same fashion as our counting method; that is, the number of unlock codes for a combination lock is not determined by a combination counting strategy. Others' general use of the term "combination" may not align with our use of the term in counting strategies. We also note the computation in a spreadsheet using $= \text{PERMUT}(40, 3)$ in a cell produces this same computed value of $59,280$. For the remainder of our examples, we will use the spreadsheet for computation work after determining the type of counting method needed.

3. A qualified applicant pool for a large hospital's eight nurse trainee positions consists of ten women and six men.
   a. How many different groups of these sixteen applicants can be formed for the eight positions?
   b. How many different groups can be selected from the pool of applicants that consist entirely of women?
   c. If all applicants are considered equally qualified, and the group of eight positions is randomly selected, what is the probability that the chosen group will consist entirely of women? (This measure could be of interest in a "discrimination" claim if the chosen pool consisted only of women.)

Answer

In this situation, we have sixteen applicants; eight are needed to form our trainee group.

   a. We recognize that we are choosing eight from our group of sixteen, and the order of arrangement of any group of eight will not make a different trainee group. This is a combination counting situation, so we compute 16 choose 8.

$$_{16}C_8 = \text{COMBIN}(16, 8) = 12,870$$

   There are well over $10,000$trainee groups that can be formed from the pool of applicants.
   b. For the selected group to be all women, we recognize that we are choosing eight from our group of ten women. We again notice that the order of arrangement of any group of eight women will not make a different trainee group; the selection order does not matter. This is a combination counting situation; we compute 10 choose 8.

$$_{10}C_8 = \text{COMBIN}(10, 8) = 45$$

   There are 45 trainee groups that can be formed only from the pool of women applicants.
   c. Assuming random selection in which any group of eight is equally likely to be chosen, the probability that the selected group will consist entirely of women is $P(\text{ALL WOMEN}) = \frac{45}{12,870} \approx 0.3497\%$,which we would consider an unusual event.

4. To win a specific Powerball lottery grand prize, a player must match all 5 white balls (order does not matter) and the red Powerball on his purchased ticket. There are 69 white balls labeled 1 to 69 and 26 red balls labeled 1 to 26.
   a. How many different tickets could be purchased for entry into this type of Powerball lottery game?
   b. Knowing there is only one winning ticket per game, what is the probability that a randomly generated ticket will be the winning ticket for the grand prize?

Answer

Notice that initially, we can see two steps in this activity of completing an entry to the Powerball game. We choose five numbers from a list of sixty-nine and then select one number from a list of twenty-six. We use our multiplication rule between the counts of the two steps.

a. For the first step, we choose any five of the sixty-nine white-ball numbers, recognizing that once a specific group of five numbers is selected, rearranging those five chosen numbers will not make a different Powerball ticket. This is a combination counting situation, so we compute ${}_{69}C_5 = 11,238,513$.For the second step, we choose any of the twenty-six red-ball numbers, with a count of 26 possible choices. (We could do a combination count of ${}_{26}C_1 = \text{COMBIN}(26,1)$ as well, but this still produces that same value of 26 and involves unnecessary computation work. Thinking as we make our counts is more valuable than computation work.) There are a total of ${}_{69}C_5 \cdot 26 \backslash 9 =$ 292,201,338 \) possible tickets that could be made for any single Powerball lottery game.

b. Assuming a random selection of numbers for the winning ticket, the probability that a randomly generated ticket will be the winning ticket is $P(\text{WINNING TICKET}) = \frac{1}{292,201,338} \approx 3.4223 \times 10^{-9}$. As we are probably all aware, winning a Powerball Lottery game by purchasing a single ticket is a highly unusual event. With these large numbers, we can also reason that we cannot create and purchase all possible tickets to guarantee winning.

Section Summary

This section on counting has been extensive. As a final review, let us list our various counting strategies.

1. Tree diagram (useful for creating and listing all possible outcomes for a situation in which multiple steps are used to produce the branches of the trees).
2. Multiplication Rule for Counting
3. Factorial Rule for Counting
4. Permutation Rule for Counting
5. Combination Rule for Counting

Knowing which counting method applies requires us to think clearly about the selection of the objects in steps and to determine if the order of selection matters to the count.

3.2.1: Counting with Indistinguishable Objects - Optional Material

Learning Objectives

- Develop and use counting methods with at least some indistinguishable objects in a situation

Counting Situation with Indistinguishable Objects

In the previous counting examples, we were dealing with counting the number of outcomes that can occur when working with distinct objects: the objects were all different from each other in the selection. We now address a situation where some of the objects are the same (indistinguishable).

Consider the possible ways we can arrange five coins, two of which are indistinguishable gold (G) coins and three are indistinguishable silver (S) coins (we compare this with our Example $3.2.1.2.1$in which all five coins were distinguishable). Our previous work demonstrated that five distinguishable coins can be arranged in $5! = 120$ ways. However, in the case of indistinguishable coins, an order of GSSGS would be counted multiple times using our factorial counting approach. Using our same reasoning and computation adjustment in developing a combination count, we adjust the size of the total ordered count of $5!$ by scaling down by the number of ways the two gold coins could be placed (there are $2!$ such ways) and the number of ways the three silver coins could be placed (there are $3!$ such ways). There are $\frac{5!}{2!\cdot 3!} = \frac{120}{12} = 10$ ways we can arrange two indistinguishable gold coins with three indistinguishable silver coins. We might list all ten possibilities to verify our computational reasoning.

In this case of just two groups of indistinguishable objects, we might notice that this computation is technically the same as our combination method ${}_5C_2 = \frac{5!}{2!\cdot(5-2)!}$. If we have two types of indistinguishable objects in our collection of five, with two being of one type, the remaining three must be of the other kind. In this situation, counting the number of ways we can select objects from among five with the order not mattering (i.e., a combination) is the same as counting the number of ways we can place two indistinguishable gold coins among five positions. The situations sound different, but the counting is the same.

Now, we can generalize beyond having two indistinguishable objects with the new counting method called the **Permutation Rule with Some Objects Indistinguishable**. In general, if there are n objects to order where all n-objects are among precisely one of the k-indistinguishable groups of n_1 alike, n_2 alike,..., n_k alike, then the number of recognizable different sequences (or permutations) of all n objects can be determined by

$$\frac{n!}{n_1!\cdot n_2!\cdots n_k!}.$$

Suppose we have ten coins, five of which are gold, two are silver, and three are copper. How many ways can these ten coins be arranged, given that each of the coins within the three metal groups is indistinguishable from the other? Using our recently developed counting strategy, we have $\frac{10!}{5!\cdot 2!\cdot 3!} = \frac{3{,}628{,}800}{120\cdot 2\cdot 6} = 2{,}520$ different distinguishable arrangements of these coins. Remember, in this calculation, we are taking all possible ordered arrangements as if we could tell the difference in the tens coins and scaling down using division by the number of arrangements each specific group can take.

? Text Exercise 3.2.1.1

Determine the number of outcomes for each of these situations.

1. Ben was born on $10/20/2004$.He wishes to use these 8 digits of his birth date to form a security code for a door lock. How many different door locks can he make from his birth date digits?

Answer

We notice that Ben has only four distinguishable digits in his 8-digit birth date, namely zeros, ones, twos, and fours. We also notice there are 4 zeros, 1 one, 2 twos, and 1 four in that date. To make an 8-digit code by different arrangements of these digits from his birth date, we count by computing $\frac{8!}{4!\cdot 1!\cdot 2!\cdot 1!} = \frac{40{,}320}{24\cdot 1\cdot 2\cdot 1} = 840$ different codes. We notice that there are not very many different possibilities. If someone knew Ben was using the digits of his birth date, a computer program could quickly produce and possibly apply 840 codes to breach his security.

2. Mikala tosses a fair coin 9 times. What is the probability that her 9 tosses will yield exactly 3 heads and 6 tails?

Answer

First, we notice we are asked a probability question with a fair coin. We use the classical method to determine this probability since each outcome is equally likely; we must determine the total number of outcomes possible and the number of those outcomes that match our desired event description. When a coin is tossed 9 times, we know by our simple multiplication rule that there are $2 \cdot 2 \cdots 2 = 2^9 = 512$ different outcomes in the ordered sample space. However, the tosses that involve 3 heads and 6 tails can happen in multiple ways, such as HHTTTHTTT or TTTHHHTTT. We must count the number of ways we can sequence 3 heads among the 9 positions (notice this is the same as asking for the number of ways we can sequence 6 tails among the 9 positions since the coin can only land either on heads or tails. There is no third indistinguishable outcome group). The three heads in any position cannot be distinguished and the six tails cannot be distinguished, so there are $\frac{9!}{3! \cdot 6!} = \frac{362,880}{6 \cdot 720} = 84$ ways to have 9 tosses that results in 3 heads and 6 tails. Mikala has the probability of $\frac{84}{512} \approx 16.41\%$ of such an outcome. Again, we note that in this situation of "two-indistinguishable groups among all n trials", we could have computed our combination count ${}_9C_3$ or ${}_9C_6$ and received the same count of 84.

3. Lanee has twelve cows to use in a study to compare three different diets, A, B, and C. Each of the diets is to be used on four of the cows. How many ways can the diets be assigned to the twelve cows?

Answer

We might first notice that a possible assignment of the diets to an ordering of the twelve cows could be:

$$\text{A B C A B C A B C A B C.}$$

A different assignment could be:

$$\text{A A B B C C C C B B A A.}$$

By these two examples, we see that a count of the number of ways in which we can assign the letters to twelve positions (representing the cows) will answer the question. There are $\frac{12!}{4! \cdot 4! \cdot 4!} = \frac{479,001,600}{24 \cdot 24 \cdot 24} = 34,650$ ways Lanee might assign the diets.

As a side note, some will call this the method of Distinguishable Permutations because we are counting the number of arrangements that can be distinguished from each other. If you have some indistinguishable objects, as we discussed above, some of the orders are not distinguishable due to some of the objects being indistinguishable. We only want to count distinguishable orderings, hence the terminology Distinguishable Permutations.

3.3: Counting and Compound Events

Learning Objectives

- Recognize distinct cases when counting
- Define compound events
- Understanding events using "or", "and", and "given"
- Develop and use the Addition Rule for counting the number of possible outcomes
- Develop and use a Multiplication Rule for counting the number of possible outcomes with compound events viewed sequentially
- Count the number of possible outcomes for various compound events

Counting: Distinct Cases

Recall Text Exercise 3.2.1b, where we were drawing two tiles in sequence without replacement from a bag consisting of 3 red, 2 white, and 1 blue tiles. We produced our sample space by constructing a tree diagram (reproduced below).

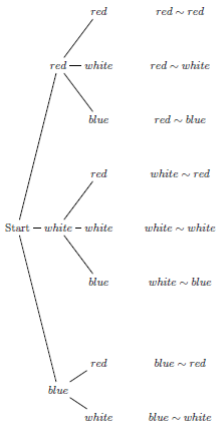


Figure 3.3.1: Tree diagram of tile combinations

This tree diagram looks different from most of the tree diagrams from the previous section; two branches end with three options, but the bottom branch only has two options. The situation and the count differed based on which branch we initially started. We had to consider two cases: case 1 of blue initially drawn and case 2 of red/white initially drawn. Either case 1 or case 2 must happen because an initial tile must be drawn, but note that both cases are exclusive. If case 1 happens, blue is the only option for the initial draw, and then there are two options for the second tile, making our count for case 1 is $1 \cdot 2 = 2$. If case 2 happens, there are two options for the initial draw and three options for the second draw, making our count for case 2 $2 \cdot 3 = 6$. We have $1 \cdot 2 + 2 \cdot 3 = 8$ possible outcomes in the sample space between the two cases.

Here is a more challenging example. Suppose a committee of ten must be formed from a group of twenty hourly employees and five managers. How many different committees can be formed if at least three managers must be on the committee? We must think carefully about how the committee is formed to count correctly. The phrase "at least three managers" tells us our selected group can have three managers or, four managers or, five managers on the committee. The committee size remains at 10 regardless of the number of managers. We can construct a table to determine the different committee compositions at the manager and hourly employee levels.

Committee Total	10	10	10
Number of Managers	3	4	5
Number of Hourly Employees	$10-3=7$	$10-4=6$	$10-5=5$

Each column of the table represents a particular committee. We recognize these committee compositions as distinct cases, one of which must happen. To count the total number of committees, we count the number of committees in each case and add them together.

$$\begin{gathered}\text{total number of committees} = \\ \text{number of committees with three managers} + \text{number of committees with four managers} \\ + \text{number of committees with five managers}\end{gathered}$$

Each case is similar in the fact that both managers and hourly employees must be selected. Our multiplication rule applies nicely since we have each group of hourly workers for each group of managers. We proceed case by case.

If the number of managers is 3, we must choose 3 of the 5 managers, order of selection does not matter, yielding ${}_5C_3$ different possibilities for filling the manager positions. The remaining 7 committee members will be chosen from the 20 hourly employees, yielding ${}_{20}C_7$ different possibilities (again, we use combinations since order does not matter). Our multiplication rule informs us that there are ${}_5C_3 \cdot {}_{20}C_7 = 10 \cdot 77,520 = 775,200$ committees possible in which three of the members are managers.

Similarly for the other two cases, there are ${}_5C_4 \cdot {}_{20}C_6 = 5 \cdot 38,760 = 193,800$ possibilities with four managers and ${}_5C_5 \cdot {}_{20}C_5 = 1 \cdot 15,504 = 15,504$ possibilities with five managers.

$$\begin{aligned}&\text{number of committees with three managers} + \text{number of committees with four managers} + \text{number of committees with five managers} \\ &= {}_5C_3 \cdot {}_{20}C_7 + {}_5C_4 \cdot {}_{20}C_6 + {}_5C_5 \cdot {}_{20}C_5 \\ &= 775,200 + 193,800 + 15,504 \\ &= 984,504\end{aligned}$$

There are $984,504$ possible committees that fit this description. From this work, we can also form some probability claims. If a committee is randomly formed given these restrictions, the probability that the committee will have five managers as members will be $\frac{15,504}{984,504} \approx 1.5748\%$ and the probability that the committee will have three managers as members will be $\frac{775,200}{984,504} \approx 78.7402\%$.

? Text Exercise 3.3.1

1. A food truck makes tacos with at least one but up to five fillings from eight ingredients: beef, fish, pork, beans, cheese, lettuce, tomatoes, and guacamole. How many ways can a taco be ordered from the food truck vendor?

Answer

We notice that we are given restrictions on the taco fillings. We must choose at least one filling for a taco while having at most five fillings. There are several distinct cases to consider, with only one taking place.

- One filling chosen
- Two fillings chosen
- Three fillings chosen
- Four fillings chosen
- Five fillings chosen

These are the only possible choices satisfying the requirements. We can determine the counts for each of these five cases and add those counts to find the total number of possibilities. Using our earlier reasoning that each of these five cases is a combination count (order of filling choices does not matter.)

$$\#\text{ of one filling} + \#\text{ of two fillings} + \#\text{ of three fillings} + \#\text{ of four fillings} + \#\text{ of five fillings}$$

$$_8C_1 + {}_8C_2 + {}_8C_3 + {}_8C_4 + {}_8C_5$$

$$8 + 28 + 56 + 70 + 56$$

$$218$$

A customer has 218 possible tacos that could be ordered. No wonder it takes some customers so long to decide what to eat.

2. A ten-person committee to investigate possible corruption in U.S. military contracts is to be formed among the 59 members of the House of Representatives Armed Services Committee and the 25 members of the Senate Committee on Armed Services. The committee must have at least five House members and at least two senators. How many committees are possible?

Answer

We notice that we are given restrictions on the committee structure. We must have at least five house members while still having at least two senators on the committee of ten. We have several distinct cases to consider, with only one taking place.

- Five House members and five Senators
- Six House members and four Senators
- Seven House members and three Senators
- Eight House members and two Senators

These are the only four possible committee structures satisfying the requirements. We can determine the counts for each of these four cases and sum those counts to find the total number of possible committees. We use similar reasoning to compute:

$$_{59}C_5 \cdot {}_{25}C_5 + {}_{59}C_6 \cdot {}_{25}C_4 + {}_{59}C_7 \cdot {}_{25}C_3 + {}_{59}C_8 \cdot {}_{25}C_2$$

$$2.66 \times 10^{11} + 5.70 \times 10^{11} + 7.85 \times 10^{11} + 6.65 \times 10^{11}$$

$$2.29 \times 10^{12}$$

There are about $2,290,000,000,000$ committees that could be formed: an astounding number of possibilities. We are glad we only had to determine the number of possible committees and were not required to list the options by constructing a tree diagram.

Multiple counting strategies come into play as we explore possible outcomes. Counting can be pretty complicated. We are on the verge of a formal addition rule that will come in handy; we will formulate it within the context of compound events.

Connecting Events Using "or", "and", and "given"

When describing multiple events, we typically use one of the following: A or B, or A and B. Using multiple events to describe or form another event is called a **compound event**.

Consider rolling a single fair die. The sample space is the set: $\{1, 2, 3, 4, 5, 6\}$.

The compound event ODD involves three outcomes: 1, 3, and 5. We understand the event ODD to occur if when we roll the die, either a 1, 3, or 5 lands face up; there are three ways for our compound event to occur. Since the die is fair, we can compute the probability using the classical method: $P(\text{ODD}) = \frac{3}{6} = \frac{1}{2}$. The complement of ODD consists of the outcomes 2, 4, and 6 which we note is the compound event EVEN; so, $\overline{\text{ODD}} = \text{EVEN}$.

The compound event PERFECT SQUARE involves just two outcomes: 1 and 4. $P(\text{PERFECT SQUARE}) = \frac{2}{6} = \frac{1}{3}$.

Consider the compound event $A = \text{EVEN}$ or PERFECT SQUARE. We refer to a single compound event A using two events connected with the word **"or."** We understand event A to occur if, when we roll the die, either EVEN occurs or PERFECT SQUARE occurs. Meaning, if 1, 2, 4, or 6 land face up, A occurs. $P(A) = P(\text{EVEN or PERFECT SQUARE}) = \frac{4}{6} = \frac{2}{3}$.

Side note: there are different ways to consider "or". Since we include the possibility that both A and B occur, our use of the word "or" is called **inclusive**. If we excluded the possibility that both events occur simultaneously, the word "or" would be called **exclusive**. Whenever we use the word "or" in this text, we are using it inclusively.

In our last example, notice that the outcome 4 was described by both EVEN and PERFECT SQUARE. When two events have outcomes in common, the events are not **mutually exclusive.**

Consider the compound event $B =$ EVEN and PERFECT SQUARE. We are again referring to a single compound event B using two events; this time, the two events are connected with the word **"and"**. We understand event B to occur if, when we roll the die, both EVEN and PERFECT SQUARE occur; meaning B will occur only when 4 lands face up. So, $P(B) = P(\text{EVEN and PERFECT SQUARE}) = \frac{1}{6}$.

For a final example, consider the event EVEN given PERFECT SQUARE which is often denoted EVEN|PERFECT SQUARE. For a third time, we refer to a single event, say event C, using two events connected with the word **"given."** When we use "given," we treat our situation as if we had additional knowledge about what is happening; that is, we assume that the die roll is guaranteed, having a probability of 1, to be a perfect square and are considering the likelihood, given this assumption, that it is even. We treat our situation **conditionally** by restricting our sample space to the event that follows the word "given." When examining the conditional event EVEN given PERFECT SQUARE, we do not consider the original sample space of $\{1, 2, 3, 4, 5, 6\}$. Instead, we only consider the perfect squares from that conditional sample space: 1 and 4. We have a "new" conditional sample space of size 2. We then ask, how many of these 2 possibilities satisfy the event EVEN? There is only one possibility of an even, 4. Now in computing the probability, we must remember that we only considered the outcomes in the event PERFECT SQUARE; so the total number of outcomes is not 6 but 2. $P(\text{EVEN}|\text{PERFECT SQUARE}) = \frac{1}{2}$.

? Text Exercise 3.3.2

Consider rolling two fair dice in sequence and the events below. For each event, identify the possible outcomes described and compute the probability.

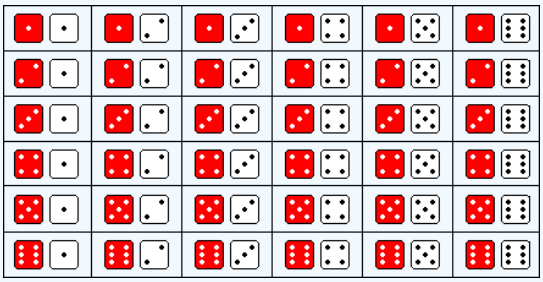


Figure 3.3.2: Sample space of rolling two standard dice

1. Event: SUM OF THE TWO VALUES IS 7

Answer

The sum of the values will be 7 for the following pairs: 1 and 6, 2 and 5, 3 and 4, 6 and 1, 5 and 2, and 4 and 3. We have $P(\text{SUM OF THE TWO VALUES IS } 7) = \frac{6}{36} = \frac{1}{6}$. Notice all these outcomes fall on the diagonal from bottom left to top right.

2. Event: EITHER DIE IS 2

Answer

We can understand the compound event EITHER DIE IS 2 as FIRST DIE IS 2 or SECOND DIE IS 2. FIRST DIE IS 2 consists of the 6 outcomes of the second row, and SECOND DIE IS 2 consists of the 6 outcomes of second column. Note that these compound events are not mutually exclusive because the outcome where both dice are 2 is in both compound events. If we count the outcomes between both compound events, we arrive at 11 outcomes in the compound event EITHER DIE IS 2. Thus $P(\text{EITHER DIE IS } 2) = \frac{11}{36}$.

3. Event: SUM OF THE TWO VALUES IS 7 and EITHER DIE IS 2

Answer

We need both SUM OF THE TWO VALUES IS 7 and EITHER DIE IS 2 to occur for our event to occur. We thus look for the overlap between the outcomes in each compound event. Only two of the outcomes from SUM OF THE TWO VALUES IS 7 have a 2 in them. Thus there are only two outcomes in the compound event SUM OF THE TWO VALUES IS 7 and EITHER DIE IS 2. Thus $P(\text{SUM OF THE TWO VALUES IS 7 and EITHER DIE IS } 2) = \frac{2}{36} = \frac{1}{18}$.

4. Event: SUM OF THE TWO VALUES IS 7|EITHER DIE IS 2

Answer

We first recall that | denotes "given." Thus our event of interest is SUM OF THE TWO VALUES IS 7 given EITHER DIE IS 2. We know that there are only 11 outcomes in the event EITHER DIE IS 2 and that only 2 of those outcomes (2 and 5, and 5 and 2) add up to 7. Thus $P(\text{SUM OF THE TWO VALUES IS 7}|\text{EITHER DIE IS } 2) = \frac{2}{11}$.

5. Event: EITHER DIE IS 2|SUM OF THE TWO VALUES IS 7

Answer

Our event of interest is EITHER DIE IS 2 given SUM OF THE TWO VALUES IS 7. From our previous exercise, we know that there are only 6 outcomes in the event SUM OF THE TWO VALUES IS 7 and that only 2 of those outcomes, 2 and 5, and 5 and 2, include 2. Thus $P(\text{EITHER DIE IS 2}|\text{SUM OF THE TWO VALUES IS 7}) = \frac{2}{6} = \frac{1}{3}$.

Terminological Consideration: Simple Events

Some statisticians and educators introduce the term simple event to help students better handle complex event descriptions. The general idea is that we can better understand or more easily compute probabilities when complex event descriptions are understood using combinations of other most basic events called simple events. The difficulty in such a presentation lies in the subjective and experiential aspects of understanding descriptions as basic and straightforward. To avoid the technical difficulty of a sufficient definition and application, we avoid the term throughout the book but, hopefully, still manage to instill the underlying concept.

Compound Events and Counting

Again, we build our intuition by counting when rolling a single fair die. When we were considering the compound event EVEN or PERFECT SQUARE, computing the probability boiled down to two counting questions: how large is the sample space and how many outcomes comprise the compound event? The first counting question has an answer of 6. Having already worked the problem, we know the answer to the second question is 4, but how does this relate to the two separate events that form our compound event? In counting the outcomes in EVEN, we arrive at 3, and the number of outcomes in PERFECT SQUARE is 2. It would appear that the number of outcomes in EVEN or PERFECT SQUARE is $3+2=5$. Our previous work shows that the answer is 4. We must have counted some outcome(s) twice, which can happen whenever the events are not mutually exclusive. Recall that the outcome 4 is both even and a perfect square but should only be counted once. The outcomes that are counted twice are in both EVEN and PERFECT SQUARE. The total count is the number of outcomes in EVEN plus the number of outcomes in PERFECT SQUARE minus the number of outcomes in EVEN and PERFECT SQUARE; namely, $3+2-1=4$.

Formally, this is known as the **Addition Rule for Counting**. Given events A and B,

$$\#\text{ of outcomes in }(A\text{ or }B)=\#\text{ of outcomes in }A+\#\text{ of outcomes in }B-\#\text{ of outcomes in }(A\text{ and }B)$$

? Text Exercise 3.3.3

Explain how we used the Addition Rule for Counting at the beginning of this section when counting tiles and committees.

Answer

We were attempting to determine the many outcomes in a sample space. Each case can be identified as a compound event, and we joined these compound events using the word "or." We added each of the counts together to get the total sum. This looks similar to the addition rule, except we never had to subtract anything since the cases were mutually exclusive; no two cases shared an outcome. This would be like subtracting 0 outcomes. The addition rule was at play the whole time.

Greater care must be taken when counting outcomes involving compound events formed using "and." We cannot mindlessly look to some formula that will always work; we must assess the situation and check that certain conditions are met. To develop our intuition regarding counting A and B, we turn our attention to a more robust but familiar example, rolling two fair dice in sequence, and analyze: EVEN FIRST and PERFECT SQUARE SECOND.

Context matters tremendously; read carefully and understand the current situation. We have dealt with the events EVEN and PERFECT SQUARE in the context of a roll of a single fair die. The sample space consists of 36 pairs of dice values; we are counting the number of outcomes with an even value for the first die and a perfect square for the second die. The multiplication rule worked well for determining the size of sample space because we could understand our outcomes as pairs of outcomes in sequence. Suppose we can understand EVEN FIRST and PERFECT SQUARE SECOND as two events in sequence, the same ideas apply. We apply the multiplication rule with the first activity, rolling the first die with an even face up (3 ways), and the second activity, rolling the second die with a perfect square face up (2 ways). We arrive at the conclusion there are $3\cdot 2=6$ outcomes in EVEN FIRST and PERFECT SQUARE SECOND. For this example, we can easily construct the outcomes to confirm our count: 2 and 1, 4 and 1, 6 and 1, 2 and 4, 4 and 4, and 6 and 4.

? Text Exercise 3.3.4

Within the context of rolling two die in sequence, determine the number of outcomes in ONE DIE IS EVEN WHILE THE OTHER IS A PERFECT SQUARE.

Answer

To save space, let EVEN $=E$ and PERFECT SQUARE $=PS$. We can understand the event as follows: (E FIRST and PS SECOND) or (PS FIRST and E SECOND). We can count the number of outcomes using both addition and multiplication rules. Applying the same logic as in the previous example and recognizing that 4 satisfies both events, yields:

$$\begin{gathered}\#\text{ in }E\text{ and }PS\\ =\#\text{ in}\\ (E\text{ FIRST and }PS\text{ SECOND})+\#\text{ in }(PS\text{ FIRST and }E\text{ SECOND})-\#\text{ in }((E\text{ FIRST and }PS\text{ SECOND})\text{ and }(PS\text{ FIRST and }E\text{ SECOND}))\\ =3\cdot 2+2\cdot 3-1\\ =6+6-1=11\end{gathered}$$

Note that it might be tempting to describe ONE DIE IS EVEN WHILE THE OTHER IS A PERFECT SQUARE as the compound event EVEN and PERFECT SQUARE. In the context of rolling two dice, we may interpret the event differently. For example, rolling a double 4 (satisfying both events), one die is even, and the other is a perfect square (one die satisfying one and the other die satisfying the other), or as long as both conditions are met between the two dice, the event occurs (a 4 with any other die value). Clarity in communication is key; put thought into the event description to protect it from ambiguity to the best of your ability.

To conclude this section and formalize one last component of counting compound events, we consider a familiar context in probability: drawing cards from a standard deck of playing cards. We draw two cards in sequence from a single deck of standard playing cards without replacing the cards drawn and examine the compound event SPADE FIRST and BLACK CARD SECOND.

Our counting procedure closely models the counting in the previous section. The phrasing of the compound event helps us readily understand our compound event as consisting of pairs of outcomes in sequence and uses the multiplication rule. There is, however, one stark contrast. In rolling two fair die and considering EVEN FIRST and PERFECT SQUARE SECOND, the value of the first die did not affect the possible values of the second die. Whatever card we draw first cannot be drawn again in our current context. There is a dependency between the events. This does not hinder our ability to count using the multiplication rule; we have run into this idea before in Text Exercise 3.2.2b. There are 13 spades and 26 black cards in a standard deck of playing cards. There are 13 ways for SPADE FIRST to happen. Since we are counting the

number of ways SPADE FIRST and BLACK CARD SECOND can happen, we count the number of ways our second card can be black given our first card was a spade. Since spades are black, only 25 black cards are left. We have the total number of outcomes $13 \cdot 25 = 325$.

We provide a **Multiplication Rule for Counting Compound Events with "and."** Given events A and B,

$$\#\text{ of outcomes in } A \text{ and } B = (\#\text{ outcomes in A}) \cdot (\#\text{ outcomes in B}|\text{A})$$

? Text Exercise 3.3.5

Within the context of drawing two cards in sequence from a single deck of standard playing cards without replacing the cards drawn, determine the number of outcomes in the event BLACK CARD FIRST and SPADE SECOND.

Answer

We are in a similar situation to the previous example; however, the order of the events has switched. We can easily understand the events as in sequence and apply the multiplication rule.

$$\#\text{ of outcomes in BLACK CARD FIRST and SPADE SECOND} = (\#\text{ outcomes in BLACK CARD FIRST}) \cdot (\#\text{ outcomes in SPADE SECOND}|\text{BLACK CARD FIRST})$$

The count in the second term depends on whether or not a spade was drawn the first time. We, therefore, need to reformulate our approach. The event BLACK CARD consists of all the spades and clubs; so we can think of the event BLACK CARD as the compound event SPADE or CLUB. Let us refer to these events as S and C, respectively. So we can understand our event as follows:

$$(\text{S or C FIRST}) \text{ and S SECOND}$$
$$(\text{S FIRST and S SECOND}) \text{ or } (\text{C FIRST and S SECOND})$$

We can thus apply both addition and multiplication rules to compute our total number of outcomes.

$$\#\text{ of outcomes in S FIRST and S SECOND} = 13 \cdot 12 = 156$$
$$\#\text{ of outcomes in C FIRST and S SECOND} = 13 \cdot 13 = 169$$
$$\#\text{ of outcomes in (S FIRST and S SECOND) and (C FIRST and S SECOND)} = 0$$

Thus the total number of outcomes is $156 + 169 - 0 = 325$.

? Text Exercise 3.3.6

A family of 5 is attending a convention on family life. The theme of this year's convention is nature and quality time. The opening banquet will have 4 door prizes related to the current theme. The door prizes, in order, are a camper, a smokeless fire pit and patio furniture, a trampoline, and a set of bicycles. Any person in attendance can win at most one prize. The family of 5 recently invested in their backyard patio with new furniture and have a trampoline but are very interested in the other two prizes. If there will be 400 people in attendance, how many ways can the door prizes be awarded so that this family gets the first and the fourth prizes?

Answer

While winning quality prizes even if you do not need them can be exciting, the question excludes the case where the family wins all four prizes. The event of interest is WINNING FIRST and LOSING SECOND and LOSING THIRD and WINNING FOURTH. There are 5 ways of winning the first prize (one of the five family members must be chosen the winner), 395 ways of not winning the second prize given a successful win of the first prize, and 394 ways of not winning the third prize given the desired outcomes of the first and second prizes, and just 4 ways to win the last prize given the outcomes of the first three prizes. There are $5 \cdot 395 \cdot 394 \cdot 4 = 3,112,600$ possibilities. This could happen in many ways, but what is the probability that it does happen? There are $400 \cdot 399 \cdot 398 \cdot 397 = 25,217,757,600$ ways the prizes could be awarded. The probability of this becoming a reality is very small. $\frac{3,112,600}{25,217,575,600} \approx 0.01234\%$. It would be highly unlikely for the family to win the first and fourth prizes.

3.4: Probability and Compound Events

Learning Objectives

- Compute probability in compound events involving "and"
- Compute probability in compound events involving "or"
- Compute probability in compound events involving "and" and "or"
- Use complements for computing the probability of compound events
- Determine probability from two-way tables
- Extend the use of probability rules to other compound events

Review and Preview Probability and Compound Events

In Chapter 3, we have discussed basic probability and counting concepts to prepare us for our future work with inferential statistics. To use the classical approach for determining the probability of an event A, we need to know the size of the sampe space, n, and we need to know that each of those outcomes in the sample space is equally likely to occur. Then, if event A can occur in x ways in that sample space, $P(A) = \frac{x}{n}$. We also discussed the empirical/experimental approach where we collect data (preferably a large data set to satisfy our Law of Large Numbers more completely) related to our situation of interest. Then, in that data set of size n, if x of the data satisfy the description of event A, we state $P(A) = \frac{x}{n}$. However, we must remember that the empirically based probability value only estimates the actual probability when the data is from a sample set. For example, in Section 2.1, we had a relative frequency table on M&M colors built on data from a package of 55 M&M's, as shown below.

Table 3.4.1: Frequencies and Relative Frequencies of Sampled M&M's

Color	Relative Frequency
Brown	$\frac{17}{55} \approx 0.309$
Red	$\frac{18}{55} \approx 0.327$
Yellow	$\frac{7}{55} \approx 0.127$
Green	$\frac{7}{55} \approx 0.127$
Blue	$\frac{2}{55} \approx 0.036$
Orange	$\frac{4}{55} \approx 0.073$

From this empirical evidence, we have probability estimates for all M&M colors. If we picked a single M&M from the population of all M&M's ever made, we would estimate, empirically, $P(\text{RED M\&M}) \approx 0.327$ and $P(\text{BLUE M\&M}) \approx 0.073$.

We have also discussed several counting rules to help us determine the size of an entire sample space and the number of outcomes matching an event description. We established various multiplication rules for counting and an addition rule for counting. Each of these rules has conditions for their use. For example, to use the general permutation rule, we had n distinct objects in which we were interested in how many ways we could select r of those objects where the order of selection mattered. Knowing the conditions allowed us to quickly determine if a given event could be counted by a specific counting method. We now extend the multiplication and addition rules of counting to similar multiplication and addition rules for probabilities of compound events.

Probability with Compound Events Involving "and"

As discussed, some event descriptions can be compound: descriptions involving multiple events. We have discussed two phrases used to combine events: "and" as well as "or." For example, in rolling two dice, the event of ROLLING A TWO ON THE FIRST DIE and ROLLING A FIVE ON THE SECOND DIE is a compound "and" description. As another example, the event of PERSON IS OVER SIXTY YEARS OLD or IS UNDER TWELVE YEARS OLD is a compound "or" description. The event of THREE MANAGERS and SEVEN HOURLY EMPLOYEES or FOUR MANAGERS and SIX HOURLY EMPLOYEES is a compound description involving both "and" as well as "or" descriptions. Although compound event descriptions may get complicated, we aim to simplify determining the probability of such compound events by developing computation methods similar to our counting methods.

We will deal first with the "and" type of event descriptions. Our previous work with probability established that with fair dice, $P(\text{TWO ON FIRST DIE}) = \frac{1}{6}$ and $P(\text{FIVE ON SECOND DIE}) = \frac{1}{6}$. Our work with the sample space of all outcomes of two fair dice showed $P(\text{TWO ON FIRST DIE and FIVE ON SECOND DIE}) = \frac{1}{36}$. We notice that the product of our two event probabilities, $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$, is the same value as the compound "and" event probability. A multiplication operation happens within probability, just as in counting. Also shown in Section 3.3, we must carefully check if the events separated by "and" depend on each other. Remember the example of the deck of cards SPADE FIRST and BLACK CARD SECOND. When the events are dependent, the count changes for the second event causing the probability to change. $P(\text{SPADE FIRST and BLACK CARD SECOND}) = \frac{13}{52} \cdot \frac{25}{51}$. We must notice if the probability of the second event changes when the first event has occurred, as in the card example. With this added condition, we still see multiplication between the two event probabilities. We have a useful general **Probability Multiplication Rule**. Given events A and B,

$$\begin{aligned} P(A \text{ and } B) &= P(A) \cdot P(B|A) \\ &= P(B) \cdot P(A|B) \end{aligned}$$

Recall from Section 3.3 that $B|A$ is read as "event B given event A has occurred." Such a probability involving the "given" condition, $P(B|A)$ is called a **Conditional Probability**.

There are events in which $P(B|A) = P(B)$ or equivalently in which $P(A|B) = P(A)$; the occurrence of one event does not affect the occurrence of the other event. For example, the results of the first die roll do not impact the second die roll. In events where $P(B|A) = P(B)$, we say that the two events A and B are **independent** of each other. But, in our example of selecting a spade card first and then a black card second, the occurrence of the first event did impact the occurrence of the second event. In events where $P(B|A) \neq P(B)$, we say that the two events A and B are **dependent** of each other. The conditional probability analysis is not needed if the

independence of the events is already known, we need only multiply the two simple event probabilities together, a **Probability Multiplication Rule for Known Independent Events**:

$$\begin{aligned} P(A \text{ and } B) &= P(A) \cdot P(B) \\ &= P(B) \cdot P(A). \end{aligned}$$

We reiterate, use of this rule requires us to know the independence of the events involved before computing our compound probability value. We see in later content that the independence of events can be required to apply specific statistical analyses. Generally, it is best to assume events are dependent and be pleasantly surprised when we can show the events are independent.

? Text Exercise 3.4.1

Answer the following compound probability questions. Recognize that the given event description is a compound event that can be broken down into multiple events.

1. We throw a fair die and randomly select a card from a standard deck. What is the probability of getting a 6 on the die and an ace on the card draw?

Answer

This is a compound "and" event description. We also notice that the two simple events are independent of each other

$$\begin{aligned} P(A \text{ and } B) &= P(\text{SIX ON DIE and ACE ON CARD DRAW} \mid \text{SIX ON DIE}) \\ &= P(\text{SIX ON DIE and ACE ON CARD DRAW}) \\ &= P(\text{SIX ON DIE}) \cdot P(\text{ACE ON CARD DRAW}) \\ &= \frac{1}{6} \cdot \frac{4}{52} \\ &= \frac{1}{78} \approx 1.2821\% \end{aligned}$$

2. We shuffle a standard deck of playing cards so the cards are randomly placed through the deck.
   a. We draw two cards. What is the probability of getting two face cards if the first card is replaced randomly in the deck before drawing the second?
   b. We draw two cards. What is the probability of getting two face cards if the first card is not replaced in the deck before drawing the second?

Answer

a. We do not want to build the sample space in this situation: there are $52 \cdot 52 = 2,704$ different outcomes. We notice that if the first card is replaced randomly in the deck before drawing the second, the two events are independent. We also notice that the deck has 12 face cards. By our multiplication rule:

$$\begin{aligned} P(A \text{ and } B) &= P(\text{FACE CARD ON FIRST DRAW and FACE CARD ON SECOND DRAW}) \\ &= P(\text{FACE ON FIRST FIRST}) \cdot P(\text{FACE ON SECOND} \mid \text{FACE CARD ON FIRST}) \\ &= P(\text{FACE ON FIRST FIRST}) \cdot P(\text{FACE ON SECOND}) \\ &\frac{12}{52} \cdot \frac{12}{52} \\ &= \frac{144}{2,704} = \frac{9}{169} \\ &\approx 5.3254\% \end{aligned}$$

We notice this is not an unusual event.

b. If the first card is not replaced in the deck before drawing the second, the second event's probability depends on the first event occurring because the number of face cards still in the deck will be down to 11. The number of cards in the entire deck will be 51 By our multiplication rule:

$$\begin{aligned} P(A \text{ and } B) &= P(\text{FACE CARD ON FIRST DRAW and FACE CARD ON SECOND DRAW}) \\ &= P(\text{FACE ON FIRST FIRST}) \cdot P(\text{FACE ON SECOND} \mid \text{FACE ON FIRST}) \\ &= \frac{12}{52} \cdot \frac{11}{51} \\ &= \frac{132}{2,652} = \frac{11}{221} \\ &\approx 4.9774\% \end{aligned}$$

Although the probability measure does not differ from the first situation drastically, it is different. The event is a little less likely to occur if the first card is not replaced; this is something we might have expected.

3. Suppose we have our bag of 55 M&M candies with the color distribution as given previously:

Table 3.4.2: Frequencies and Relative Frequencies of Sampled M&M's

Color	Relative Frequency
Brown	$\frac{17}{55}$
Red	$\frac{18}{55}$
Yellow	$\frac{7}{55}$
Green	$\frac{7}{55}$
Blue	$\frac{2}{55}$

Color	Relative Frequency
Orange	$\frac{4}{55}$

We randomly grab four candies, one at a time, without replacement. What is the probability that we have a red candy first, a green candy second, a brown candy third, and a brown candy fourth?

Answer

These events are not independent. For example, the probability of getting a green candy second will vary depending on whether we had a red candy first. By our multiplication rule:

$$\begin{aligned} P\left(\text{RED } 1^{st} \text{ and GREEN } 2^{nd} \text{ and BROWN } 3^{rd} \text{ and BROWN } 4^{th}\right) &= P\left(\text{RED}\right) \cdot P\left(\text{GREEN} \mid \text{RED } 1^{st}\right) \cdot P\left(\text{BROWN} \mid \text{RED and GREEN previously}\right) \\ &\quad \cdot P\left(\text{BROWN} \mid \text{RED and GREEN and BROWN previously}\right) \\ &= \frac{18}{55} \cdot \frac{7}{54} \cdot \frac{17}{53} \cdot \frac{16}{52} \\ &= \frac{34,272}{8,185,320} = \frac{476}{113,685} \\ &\approx 0.4187\% \end{aligned}$$

4. Senior citizens make up about 18% of the U.S. adult population, according to the Pew Research Center website.
   a. What is the probability of randomly selecting two U.S. adults who are both senior citizens?
   b. What is the probability of randomly selecting two U.S. adults, the first not a senior citizen and the second is a senior citizen?
   c. What is the probability of randomly selecting three U.S. adults such that the first two are not senior citizens and the third is a senior citizen

Answer

a. As the U.S. population of senior citizens is very large, we will assume independence in the events, realizing our probability measures are approximations. By our multiplication rule:

$$\begin{aligned} P\left(A \text{ and } B\right) &= P\left(\text{SENIOR CITIZEN and SENIOR CITIZEN}\right) \\ &= P\left(\text{SENIOR CITIZEN}\right) \cdot P\left(\text{SENIOR CITIZEN}\right) \\ &= 18\% \cdot 18\% \\ &\approx 3.24\% \end{aligned}$$

We do emphasize that assumption of independence is not always reasonable, but is used at times when dealing with large population/sample sizes as probability measures do not get impacted significantly in value. We can see this in an example of comparison on, say, a dependence measure of $\frac{2456783}{34595200} \cdot \frac{2456782}{34595199} \approx 0.504314831996\%$ versus assumed independence measure of $\frac{2456783}{34595200} \cdot \frac{2456783}{34595200} \approx 0.50431502269\%$ Although the probability values are technically different, the practical interpretation for real-life application would not usually have practical meaning in the difference. We do note that we reflect carefully before assuming independence as it can lead to drastic consequences.

b. Again, as the U.S. population is very large, we will assume independence in the events. By our complement rule, we note that about 82% of the U.S. adult population is less than 65 years old. By our multiplication rule:

$$\begin{aligned} P\left(\text{NOT A SENIOR CITIZEN and SENIOR CITIZEN}\right) &= P\left(\text{NOT A SENIOR CITIZEN}\right) \cdot P\left(\text{SENIOR CITIZEN}\right) \\ &= 82\% \cdot 18\% \\ &\approx 14.76\% \end{aligned}$$

c. Again, as the U.S. population is very large, we will assume independence in the events. By our complement rule, we note that about 82% of the U.S. adult population is less than 65 years old. By our multiplication rule:

$$\begin{aligned} P\left(\text{NOT A SENIOR CITIZEN and NOT A SENIOR CITIZEN and SENIOR CITIZEN}\right) &= 82\% \cdot 82\% \cdot 18\% \\ &\approx 12.10\% \end{aligned}$$

Probability with Compound Events Involving "or"

Now that we have handled probability measures on events separated by "and," we turn to probability and events separated by "or." For example, we might ask, "In rolling a single fair die, what is the probability of rolling a two or a three?" In the case of our M&M's, we might ask, "What is the probability of the next M&M we randomly remove from a bag being a blue or an orange M&M?" As discussed in Section 3.3, the "or" compound description is tied to addition. As a reminder, our developed Addition Rule for Counting was given as

$$\#\text{ of outcomes in } (A \text{ or } B) = \#\text{ of outcomes in } A + \#\text{ of outcomes in } B - \#\text{ of outcomes in } (A \text{ and } B)$$

Our developed rule reminds us to check for the outcomes that matched the descriptions of both events and to subtract the set of the twice-counted outcomes.

Similar to how the Multiplication Rule for Counting extends to the Multiplication Rule for Probability, the Addition Rule for Counting will also extend to an **Addition Rule for Probability**.

$$P\left(A \text{ or } B\right) = P\left(A\right) + P\left(B\right) - P\left(A \text{ and } B\right)$$

In answering the question, "In rolling a single fair die, what is the probability of rolling a two or a three?", we can apply this addition rule to produce

$$
\begin{aligned}
&P(\text{ROLLING A TWO or ROLLING A THREE}) \\
&= P(\text{ROLLING A TWO}) + P(\text{ROLLING A THREE}) - P(\text{ROLLING A TWO and ROLLING A THREE}) \\
&= \frac{1}{6} + \frac{1}{6} - \frac{0}{6} \\
&= \frac{2}{6} = \frac{1}{3} \approx 33.3333\%
\end{aligned}
$$

Since the sample space of rolling a single fair die is small in size, one could likely answer this question much faster by knowing that only two of the six outcomes in the sample space meet the description of being "EITHER A TWO or A THREE. But as we see in the text exercises, this addition rule can be very useful when the sample spaces are large.

? Text Exercise 3.4.2

Answer the following compound probability questions. Recognize that the given event description is a compound event that can be broken down into multiple events.

1. A standard deck of playing cards is well-shuffled for randomness. Determine the following probabilities about a single card draw.
   a. Find the probability that a randomly drawn card is a king or a queen.
   b. Find the probability that a randomly drawn card is black or an ace.
   c. Find the probability that a randomly drawn card is a spade or a face card.

Answer

a. This is a compound "or" event description, as the event of interest is a king or queen card.

$$
\begin{aligned}
&P(\text{CARD IS A KING or CARD IS A QUEEN}) \\
&= P(\text{KING}) + P(\text{QUEEN}) - P(\text{KING and QUEEN}) \\
&= \frac{4}{52} + \frac{4}{52} - \frac{0}{52} \\
&= \frac{8}{52} = \frac{2}{13} \approx 15.3846\%
\end{aligned}
$$

We notice that, because $P(\text{KING and QUEEN}) = 0\%$, these two events are mutually exclusive. There is a little over 15% probability that a random card draw will produce a king or a queen. Although not highly probable, we would not consider this outcome unusual.

b. This is a compound "or" event description, as the event of interest is a black card or an ace.

$$
\begin{aligned}
&P(\text{CARD IS BLACK or CARD IS AN ACE}) \\
&= P(\text{BLACK}) + P(\text{ACE}) - P(\text{BLACK and ACE}) \\
&= \frac{26}{52} + \frac{4}{52} - \frac{2}{52} \\
&= \frac{28}{52} = \frac{7}{13} \approx 53.8462\%
\end{aligned}
$$

We notice that, because $P(\text{BLACK and ACE}) \neq 0\%$, these two events are not mutually exclusive. There is almost a 54% chance that a random card draw will produce a black or an ace card; such an outcome is reasonably probable.

c. This is a compound "or" event description, as the event of interest is a card that is a spade or a face card. We note "FACE CARD" is also an "or" event since "FACE CARD" means "KING or QUEEN or JACK". Sometimes, event descriptions can be rephrased to make a compound event more evident as an "and" or an "or" type event

$$
\begin{aligned}
&P(\text{CARD IS SPADE or FACE CARD}) \\
&= P(\text{SPADE}) + (P(\text{KING}) + P(\text{QUEEN}) + P(\text{JACK})) - P(\text{SPADE and FACE CARD}) \\
&= \frac{13}{52} + \left(\frac{4}{52} + \frac{4}{52} + \frac{4}{52}\right) - \frac{3}{52} \\
&= \frac{22}{52} \\
&= \frac{11}{26} \approx 42.3077\%
\end{aligned}
$$

We again notice, because $P(\text{SPADE and FACE CARD}) \neq 0\%$, these two events are not mutually exclusive. There is a little over 42% probability that a random card draw will produce a spade or a face card.

2. An online clothing store has a liberal return policy since customers cannot try on the items before purchasing. From survey data gathered in the return process from their customers, 21% of all purchased items are returned due to the item being too small/tight, while 4% are returned due to the item being too big/loose. If a customer-purchased item is randomly selected from all purchases, what is the probability that the item will be returned due to being too big/loose or too small/tight?

Answer

This is a compound "or" event description as the purchased item must be returned due to being too big/loose or too small/tight. We notice that these two events are mutually exclusive; we assume that a clothing item cannot simultaneously be too big and too small

$$
\begin{aligned}
&P(\text{TOO BIG/LOOSE or TOO SMALL/TIGHT}) \\
&= P(\text{TOO BIG/LOOSE}) + P(\text{TOO SMALL/TIGHT}) - P(\text{TOO BIG/LOOSE and TOO SMALL/TIGHT}) \\
&= 21\% + 4\% - 0\% = 25\%
\end{aligned}
$$

Twenty-five percent of all customer-purchased items are returned due to being either too big/loose or too small/tight. Such a significant return rate for these two issues will likely require the store to find ways to help customers find a better fit for their orders.

3. We return to our bag of 55 M&M candies.

Table 3.4.3: Frequencies and Relative Frequencies of Sampled M&M's

Color	Relative Frequency
Brown	$\frac{17}{55}$
Red	$\frac{18}{55}$
Yellow	$\frac{7}{55}$
Green	$\frac{7}{55}$
Blue	$\frac{2}{55}$
Orange	$\frac{4}{55}$

a. What is the probability of randomly drawing one candy of a primary color, red, yellow, or blue?
b. What is the probability of randomly drawing one candy that is not of one of those primary colors?

Answer

a. By our addition rule and noticing the events are mutually exclusive, we compute

$$\begin{aligned} P(\text{PRIMARY COLOR M\&M}) &= P(\text{RED M\&M or YELLOW M\&M or BLUE M\&M}) \\ &= P(\text{RED}) + P(\text{YELLOW}) + P(\text{YELLOW}) \\ &= \frac{18}{55} + \frac{7}{55} + \frac{2}{55} \\ &= \frac{27}{55} \approx 49.0909\%. \end{aligned}$$

b. We could apply our addition rule again relative to the colors of brown, green, and orange. But we notice this is also a complement to the previous event.

$$\begin{aligned} P(\text{NOT PRIMARY COLOR M\&M}) &= 1 - P(\text{PRIMARY COLOR M\&M}) \\ &\approx 1 - 0.490909 \\ &= .509091 = 50.9091\% \end{aligned}$$

We notice that the probability of randomly selecting a primary color M&M is practically the same as getting a non-primary color. Stated equivalently, the proportion of M&M's is approximately 50% for each color grouping.

Extended Concepts on Compound Events

As mentioned, we sometimes have event descriptions that include multiple compounding actions. Thankfully, we do not need any extra computation rules for these. However, we do have to read very carefully to properly understand the compound event descriptions and ask for clarification if needed.

We must clarify a small detail about our "AND" rule. In our previous examples using compound "AND" event descriptions, we worked with descriptions involving "sequential trials" of two or more trials: tossing two dice, drawing four M&M candies, and randomly selecting two people. What if we are describing an event with the word "and" but in a single trial: Selecting a single M&M that is both red and brown, drawing a single card that is red and a face card? We use a compound "and," but does our multiplication rule apply in a single trial event? We can often reflect on our sample space and naturally handle the "and" within a single trial. Our general multiplication rule still works, but the two events separated by "and" will often be dependent in a single trial situation. We must carefully consider the conditional probability on the second described event as we apply the multiplication rule.

Let us first examine the probability of selecting a single M&M that is both red and brown from our bag of 55 M&M's. Based on the given information (and our experience eating M&M's), we can reason that there are no M&M's that meet that description, so the probability is $\frac{0}{55} = 0\%$. By reflecting on the sample space, we determine the probability value; no special probability rule is necessary, but let us see that the multiplication rule still works.

$$\begin{aligned} P(\text{RED and BROWN}) &= P(\text{RED}) \cdot P(\text{BROWN} \mid \text{RED}) \\ &= \frac{18}{55} \cdot \frac{0}{18} \\ &= \frac{\cancel{18}}{55} \cdot \frac{0}{\cancel{18}} \\ &= \frac{0}{55} = 0\% \end{aligned}$$

We can see that in handling the conditional probability of selecting a single M&M that is BROWN given that we are restricted to the RED M&M's produces the probability value $\frac{0}{18}$; none of the red M&M's are brown. Our multiplication rule for "AND" still works for even single trial cases. It is acceptable and encouraged to reflect on the sample space and not use our multiplication rule to determine such probability values. We should use sound reasoning, not blind use of formulas, to determine measures.

To see another example, find the probability of drawing a single card that is both red and a face card. By thinking about the sample space of a deck of cards, one might quickly reason that there are six cards in the deck of fifty-two that are red face cards. The probability is $\frac{6}{52} = \frac{3}{26} \approx 11.5385\%$. Let us check that our multiplication rule produces this value as well.

$$\begin{aligned} P(\text{RED and FACE CARD}) &= P(\text{RED}) \cdot P(\text{FACE CARD} \mid \text{RED}) \\ &= \frac{26}{52} \cdot \frac{6}{26} \\ &= \frac{\cancel{26}}{52} \cdot \frac{6}{\cancel{26}} \\ &= \frac{6}{52} \approx 11.5385\% \end{aligned}$$

We can see that in handling the conditional probability of selecting a single FACE CARD given that we are restricted to the RED cards produces the probability value $\frac{6}{52}$ since six of the red cards are face cards. Using conditional probability with the multiplication rule can be handy in more challenging event descriptions.

Let's try a few single trial "and" compound event exercises.

? Text Exercise 3.4.3

Answer the following single-trial compound probability questions. Recognize that the given event description is a compound event that can be broken down into multiple events.

1. We draw a single card from a well-shuffled standard deck of playing cards. What is the probability of getting a black ace?

Answer

This is a compound "and" event description as the card must be both black and an ace.

$$\begin{aligned} P(\text{BLACK CARD and ACE CARD}) &= P(\text{BLACK CARD}) \cdot P(\text{ACE CARD} \mid \text{BLACK CARD}) \\ &= \frac{26}{52} \cdot \frac{2}{26} \\ &= \frac{2}{52} = \frac{1}{26} \\ &\approx 3.8462\% \end{aligned}$$

We could have answered this by reflecting on the sample space of the 52 outcomes, realizing there are 2 black aces, but our multiplication rule does work.

2. Based on $2023-2024$ information from the U.S. Center for Disease Control website, about 70% of senior citizens (65 years and older) in the U.S. get the flu vaccine, whereas about 40 of those adults under 65 years old get vaccinated. Senior citizens make up about 18% of the U.S. adult population, according to the Pew Research Center website.
   a. What is the probability of randomly selecting one U.S. adult who is a senior citizen and has had the flu shot?
   b. What is the probability of randomly selecting one U.S. adult who is not a senior citizen and has not had the flu shot?

Answer

a. We can use our multiplication rule.

$$\begin{aligned} P(\text{SENIOR CITIZEN and FLU SHOT}) &= P(\text{SENIOR CITIZEN}) \cdot P(\text{FLU SHOT} \mid \text{SENIOR CITIZEN}) \\ &= 18\% \cdot 70\% \\ &= 0.18 \cdot 0.70 \\ &= 0.126 \\ &= 12.6\% \end{aligned}$$

In $2023-2024$,about $12.6\%$of the U.S. adult population consisted of senior citizens who had taken the flu shot.

b. We can use our multiplication rule.

$$\begin{aligned} P(\text{NOT A SENIOR CITIZEN and NO FLU SHOT}) &= P(\text{NOT A SENIOR CITIZEN}) \cdot P(\text{NO FLU SHOT} \mid \text{NOT A SENIOR CITIZEN}) \\ &= 82\% \cdot 60\% \\ &= 49.2\% \end{aligned}$$

In $2023-2024$,about $49.2\%$of the U.S. adult population consisted of non-senior citizens who had not taken the flu shot. Notice that this tells health officials about the demographics of those who did not have the flu shot.

Now, we focus on investigating our event descriptions with multiple compound events. No new probability calculation rules are necessary; we must apply our current understanding to slightly new contexts. For example, suppose we wish to find the probability of first throwing a fair die with a result of an even number and then drawing a card from a standard deck of playing cards that is either a red or a face card. This compound event description involves both "and" and "or." Reading carefully, we break down the description: a first event involving throwing a fair die, then a second event involving drawing a card. We notice these two events are independent. In the card draw event, we note that drawing a red card and a face card are not mutually exclusive. We have

$$
\begin{aligned}
&P(\text{DIE TOSS OF EVEN and } [\text{CARD DRAW OF RED or FACE}]) \\
&= P(\text{DIE TOSS OF EVEN}) \cdot P(\text{CARD DRAW RED or FACE}) \\
&\qquad = P(\text{DIE TOSS OF EVEN}) \\
&\cdot (P(\text{CARD DRAW RED}) + P(\text{CARD DRAW FACE}) - P(\text{CARD DRAW RED and FACE})) \\
&= \frac{3}{6} \cdot \left(\frac{26}{52} + \frac{12}{52} - \frac{6}{52} \right) \\
&= \frac{1}{2} \cdot \left(\frac{32}{52} \right) \\
&= \frac{1}{2} \cdot \frac{8}{13} = \frac{4}{13} \approx 30.7692\%.
\end{aligned}
$$

? Text Exercise 3.4.4

Determine the probabilities of each of the following.

1. We draw five cards from a well-shuffled standard deck of playing cards. What is the probability of getting a flush (all cards share the same suit)?

Answer

This is a compound "and" event description, as the five cards must all be of one suit. We note that the suit does not matter, so the first card drawn can be of any suit; the remaining cards must match the first. The events are not independent.

$$
\begin{aligned}
&P(\text{FLUSH}) \\
&= P\left(\text{ANY SUIT CARD } 1^{st} \text{ and SAME SUIT CARD } 2^{nd} \text{ and SAME SUIT CARD } 3^{rd} \text{ and SAME SUIT CARD } 4^{th} \text{ and SAME SUIT CARD } 5^{th}\right) \\
&= P\left(\text{ANY SUIT CARD } 1^{st}\right) \cdot P\left(\text{SAME SUIT CARD } 2^{nd} | 1^{st} \text{ CARD SUIT}\right) \cdot P\left(\text{SAME SUIT CARD } 3^{rd} | 1^{st}, 2^{nd} \text{ CARD SUIT}\right) \\
&\qquad \cdot P\left(\text{SAME SUIT CARD } 4^{th} | 1^{st}, 2^{nd}, 3^{rd} \text{ CARD SUIT}\right) \cdot P\left(\text{SAME SUIT CARD } 5^{th} | 1^{st}, 2^{nd}, 3^{rd}, 4^{th} \text{ CARD SUIT}\right) \\
&= \frac{52}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48} \\
&= \frac{33}{16660} \approx 0.1981\%
\end{aligned}
$$

It should not be surprising that flush of five cards from a well-shuffled deck is a very unlikely event.

2. If we toss a fair die and then flip a fair coin, what is the probability that we get either a 6 on the die or a head on the coin (or both)?

Answer

We approach this in two different ways. The first may initially sound easier or more natural, but the second brings us different, valuable insights. Accurate rephrasing of a given situation can make the probability calculation different yet produce the same results.

a. The basic events are a toss of a fair die followed by a flip of a fair coin. Our event description is of the "or" type. Since we can toss a 6 and get a head on the coin in one execution of the situation, we note that our events are not mutually exclusive. By our addition rule, we obtain the following.

$$
\begin{aligned}
&P(6 \text{ ON DIE or HEAD ON COIN}) \\
&= P(6 \text{ ON DIE}) + P(\text{HEAD ON COIN}) - P(6 \text{ ON DIE and HEAD ON COIN}) \\
&= \frac{1}{6} + \frac{1}{2} - \left(\frac{1}{6} \cdot \frac{1}{2} \right) \\
&= \frac{7}{12} \approx 58.3333\%
\end{aligned}
$$

b. By examining complement descriptions first, the complement of "either a 6 on the die or a head on the coin (or both)" is "not getting a 6 on the die and also not a head on the coin". Now, we can use our complement and multiplication rule.

$$
\begin{aligned}
&P(6 \text{ ON DIE or HEAD ON COIN}) \\
&= 1 - (P(\text{NOT A 6 ON DIE and NOT A HEAD ON COIN})) \\
&= 1 - (P(\text{NOT A 6 ON DIE}) \cdot P(\text{NOT A HEAD ON COIN})) \\
&= 1 - \left(\frac{5}{6} \cdot \frac{1}{2} \right) \\
&= 1 - \frac{5}{12} = \frac{7}{12} \approx 58.3333\%
\end{aligned}
$$

We do notice we get the same answer although computed through a very different approach.

3. If we roll a fair die three times, what is the probability that one or more throws will come up with a 1?

Answer

On the surface, this sounds easy, but that changes once we think about the possible ways to have one or more of the three approaches produce a 1. We could have the first toss produce a 1 while the other two do not, or we can have the second produce a 1 and the other two not, or we can have the first two tosses both produce a 1 and the third not, or many other possibilities. We must determine many probabilities to use our rule with "or." However, there is an easier way. Notice the complement of "at least one of the throws will be a 1" is given by "none of the three throws produce a 1."

We obtain the following by our complement and multiplication rule, utilizing the independence of the die throws.

$$
\begin{aligned}
&P(\text{AT LEAST ONE - 1 AMONG THREE TOSSES}) \\
&= 1 - P(\text{NOT 1 ON FIRST DIE and NOT 1 ON SECOND DIE and NOT 1 ON THIRD DIE}) \\
&= 1 - [P(\text{NOT 1 ON FIRST DIE}) \cdot P(\text{NOT 1 ON SECOND DIE}) \cdot P(\text{NOT 1 ON THIRD DIE})] \\
&= 1 - \left(\frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6}\right) \\
&= 1 - \frac{125}{216} = \frac{91}{216} \approx 42.1296\%
\end{aligned}
$$

4. Based on $2023 - 2024$ information from the U.S. Center for Disease Control website, about 70% of senior citizens (65 years and older) in the U.S. get the flu vaccine, whereas about 40 of those adults under 65 years old get vaccinated. Senior citizens make up about 18% of the U.S. adult population, according to the Pew Research Center website. What is the probability of randomly selecting two U.S. adults who are both senior citizens and recipients of the flu vaccine?

Answer

Due to the large population size, we assume the selection of individuals is independent. We believe that if the first selected adult is a senior citizen, the second selected adult still has a 18% chance of being a senior citizen; likewise, we assume that the probability that the second person chosen is vaccinated is the same as the first person. We now use our multiplication rule.

$$
\begin{aligned}
&P([\text{FIRST IS A SENIOR CITIZEN and FIRST HAS FLU SHOT}] \text{ and } [\ slightSECOND\ IS\ A\ SENIOR\ CITIZEN \text{ and SECOND HAS FLU SHOT}]) \\
&= [P(\text{FIRST IS A SENIOR CITIZEN}) \cdot P(\text{FIRST HAS FLU SHOT} \mid \text{FIRST IS A SENIOR CITIZEN})] \\
&\cdot [P(\text{SECOND IS A SENIOR CITIZEN}) \cdot P(\text{SECOND HAS FLU SHOT} \mid \text{SECOND IS A SENIOR CITIZEN})] \\
&= (18\% \cdot 70\%) \cdot (18\% \cdot 70\%) \\
&\approx 1.5876\%
\end{aligned}
$$

So in regard to this $2023 - 2024$data, it would be unusual for us to randomly select two U.S adults both of whom were flu vaccinated senior citizens.

Two-Way Tables and Probability

Sometimes, data frequencies for a collected single variable are **disaggregated**, separated into mutually exclusive subgroups. For example, quiz data frequencies for a class may be separated by the frequency of those who passed the quiz versus those who did not. Alternatively, perhaps the data is separated into those students who are in extra-curricular school activities and those who are not. If done well, this allows a comparison between subgroups.

Let us go one step further in this discussion: disaggregating data in two ways. These results are displayed in a **Two-Way Table** or a **Contingency Table**. We take the quiz data of a class of 20 students first given in Section 2.4 and separate by passing, 6 or above, and non-passing scores, below 6, then by student involvement in extra-curricular school activities resulting in the following two-way table.

Disaggregation	Passed Quiz	Failed Quiz
Not Involved in Extra-Curricular School Activities	7	1
Involved in Extra-Curricular School Activities	9	3

When reading the table, we must pay attention to the row and column headings. The value 3 in the table implies that 3 of the 20 students were involved in extra-curricular activities and failed the quiz. We often include row and column totals for each disaggregation when working with two-way tables.

Disaggregation	Passed Quiz	Failed Quiz	Row Totals
Not Involved in Extra-Curricular School Activities	7	1	8
Involved in Extra-Curricular School Activities	9	3	12
Column Totals	16	4	20

We can see a total of 20 student data points. Of those, there were 16 that passed the quiz, while there were 8 not involved in extra-curricular activities.

We can now answer several probability/proportion questions concerning these results. What is the probability that a randomly selected student from the class was involved in extra-curricular school activities? By our classical probability approach, $P(\text{IN EXTRA CURRICULAR}) = \frac{12}{20} = \frac{3}{5} = 60\%$.

What is the probability that a randomly selected student from the class failed the quiz and was not involved in extra-curricular activities? Since randomly selecting a student from the twenty, $P(\text{FAILED QUIZ and NOT IN EXTRA CURRICULAR}) = \frac{1}{20} = 5\%$. This is because the table shows only 1 student in both the FAILED QUIZ column as well as the NOT IN EXTRA CURRICULAR row. We could also apply our multiplication rule, noting the events are not independent.

$$
\begin{aligned}
&P(\text{FAILED QUIZ and NOT IN EXTRA CURRICULAR}) \\
&= P(\text{FAILED QUIZ}) \cdot P(\text{NOT IN EXTRA CURRICULAR} \mid \text{FAILED QUIZ}) \\
&= \frac{4}{20} \cdot \frac{1}{4} = \frac{1}{20} = 5\%
\end{aligned}
$$

In this situation, the first approach is much easier due to the table information.

Text Exercise 3.4.5

Answer the following probability questions about a given two-way table.

1. A polygraph device is being studied for its accuracy in lie detection. A group of 200 randomly drawn participants were divided into two groups: those instructed not to lie and those who were supposed to lie. The technician running the device did not know each participant's group. Each participant was tested with the device, and a positive or negative detection of a lie was returned. The results are given in the following contingency table.

Disaggregation	Positive Lie Detection Reading	Negative Lie Detection Reading
Participant Did Not Lie	9	71
Participant Did Lie	85	35

As an example, 71 of the 200 participants did not lie, and the detector did not sense a lie occurring.

a. What is the probability that a randomly selected participant had a positive lie detection reading?
b. What proportion of the participants had a false positive reading (the device detected a lie, but the participant did not lie)?
c. What is the probability of randomly selecting two participants in sequence so that both had a positive lie detection reading?
d. What is the probability of randomly selecting a participant who lied but had a negative detection reading or who did not lie but had a positive detection reading? In other words, for what proportion of the trials was the device inaccurate?
e. What is the probability of randomly selecting two participants in sequence so that the first lied with a positive detection reading and the second did not lie with a negative detection reading?

Answer

We first add column and row totals to our table.

Disaggregation	Positive Lie Detection Reading	Negative Lie Detection Reading	Row Totals
Participant Did Not Lie	9	71	80
Participant Did Lie	85	35	120
Column Totals	94	106	200

Now, we answer the probability/proportion questions using appropriate probability rules.

a. From the two-way table, we notice there were 94 participants that had positive lie detection readings. So, $P(\text{POSITIVE DETECTION}) = \frac{94}{200} = \frac{47}{100} = 47\%$.
b. From the table, we can notice there were 9 participants had a false positive reading. So, we can quickly produce results of $P(\text{FALSE POSITIVE DETECTION}) = \frac{9}{200} = 4.5\%$.

   We instead might have noticed this was an "and" compound event: DID NOT LIE and POSITIVE LIE DETECTION. Therefore, it would also have been appropriate (though more cumbersome) to measure the proportion using our general multiplication rule.

$$\begin{aligned} &P\left(\text{DID NOT LIE and POSITIVE LIE DETECTION}\right) \\ &= P\left(\text{DID NOT LIE}\right) \cdot P\left(\text{POSITIVE LIE DETECTION} \mid \text{DID NOT LIE}\right) \\ &= \frac{80}{200} \cdot \frac{9}{80} \\ &= \frac{9}{200} = 4.5\% \end{aligned}$$

   We mention two items in reflection of this exercise. First, we notice how the table is used for a conditional probability such as $P(\text{POSITIVE LIE DETECTION} \mid \text{DID NOT LIE}) = \frac{9}{80}$. The given condition of DID NOT LIE requires us to use only the information in the table row of "Participants Did Not Lie."

   This exercise demonstrates that multiple approaches might be taken to a probability/proportion question, but there is only one correct answer. What is important is to have sound reasoning in our approach. After that, experience is the best teacher in finding the easiest approaches without making mistakes in our reasoning.

c. We notice the event description can be re-worded as FIRST PARTICIPANT HAD POSITIVE LIE DETECTION and SECOND PARTICIPANT HAD POSITIVE LIE DETECTION, so we use the general multiplication rule.

$$\begin{aligned} &P\left(\text{FIRST PARTICIPANT HAD POSITIVE LIE DETECTION and SECOND PARTICIPANT HAD POSITIVE LIE DETECTION}\right) \\ &= P\left(\text{FIRST HAD POSITIVE DETECTION}\right) \cdot P\left(\text{SECOND HAD POSITIVE DETECTION} \mid \text{FIRST HAD POSTIVE DETECTION}\right) \\ &= \frac{94}{200} \cdot \frac{93}{199} \\ &= \frac{8,742}{39,800} = \frac{4371}{19,900} \approx 21.9648\% \end{aligned}$$

   We notice how these sequential events are not independent since we cannot pick the same participant twice.

d. We notice the event description appears to primarily be an "or" compound event. That is, our event of interest requires selection of one participant that LIED WITH NEGATIVE DETECTION or DID NOT LIE WITH POSITIVE LIE DETECTION, so we use the general addition rule noting these two events are mutually exclusive.

$$\begin{aligned} &P\left(\text{LIED WITH NEGATIVE DETECTION or DID NOT LIE WITH POSITIVE LIE DETECTION}\right) \\ &= P\left(\text{LIED WITH NEGATIVE DETECTION}\right) + P\left(\text{DID NOT LIE WITH POSITIVE LIE DETECTION}\right) \\ &= \frac{35}{200} + \frac{9}{200} \\ &= \frac{44}{200} = \frac{11}{50} = 22\% \end{aligned}$$

We observe that a test with a 22% inaccuracy rate is not very dependable. Also, we excluded the subtraction in the general addition rule for possible double counting of LIED WITH NEGATIVE DETECTION and DID NOT LIE WITH POSITIVE LIE DETECTION since we recognized the events were mutually exclusive.

e. The event description is an "and" compound type as we want sequential selection of two participants with FIRST PARTICIPANT LIED WITH POSITIVE LIE DETECTION and SECOND PARTICIPANT DID NOT LIE WITH NEGATIVE LIE DETECTION. We use the general multiplication rule.

$$\begin{aligned} &P\left(1^{st}\text{ PARTICIPANT LIED WITH POSITIVE LIE DETECTION and }2^{nd}\text{ PARTICIPANT DID NOT LIE WITH NEGATIVE LIE DETECTION}\right) \\ &\qquad = P\left(1^{st}\text{ PARTICIPANT LIED WITH POSITIVE DETECTION}\right) \\ &\qquad\quad \cdot P\left(2^{nd}\text{ PARTICIPANT DID NOT LIE WITH NEGATIVE DETECTION }|1^{st}\text{ PARTICIPANT LIED WITH POSTIVE DETECTION}\right) \\ &= \frac{85}{200} \cdot \frac{71}{199} \\ &= \frac{6,035}{39,800} = \frac{1,207}{7,960} = 15.1633\% \end{aligned}$$

It would not be unusual, since the probability value is above 5%.

CHAPTER OVERVIEW

4: Probability Distributions

4.1: Random Variables

Learning Objectives

- Define and construct random variables
- Establish that the sum of probabilities across all values of a random variable is 1
- Define and distinguish discrete random variables from continuous random variables
- Introduce the discrete uniform distribution

Section 4.1 Excel File (contains all of the data sets for this section)

Review and Preview

In inferential statistics, we seek to understand a population by studying a sample randomly selected from it. In particular, we are trying to estimate the value of a population parameter based on our study of a random sample, including many sample statistics. Hopefully, at this point, we recognize that random sampling is a random experiment; when random sampling is conducted, the outcome is a particular sample, and sample statistics are values computed from the outcome. The distinction between the outcome of random sampling (the sample) and the values that describe the outcome (statistics) is essential to remember. We are interested in the likelihood that our sample is representative of the population. To determine this likelihood, we consider all possible values that sample statistics may take on and the probabilities of such values occurring. As such, we can understand sample statistics as random variables.

It is crucial to grasp that, when we view a sample statistic as a random variable, we are not just seeing a number. We are exploring a method that can generate a value from any random sample, and diving into the realm of all the possible values that could be produced. This shift in perspective is not just a technicality, but a fundamental concept that necessitates understanding.

Random Variables

A **random variable** is a quantitative variable that assigns a number to each outcome in the sample space of a given random experiment. We generally denote random variables using capital letters, like X, and the particular values that they take on (the values that are assigned to the outcomes of a random experiment) with the same letter but lowercase and with indices: x_1, x_2, x_3, x_4,..., x_n (if there are n possible values). As we indicated above, we are most interested in connecting the values or a random variable with the probability that they occur. The values of the random variable, together with their probabilities, form the **probability distribution** of the random variable. In general, our interest lies in the probability distributions of random variables, and as you will see, we have studied some of them already.

Consider a familiar example of a random experiment: rolling two fair dice. There are many different ways in which we could construct a random variable. One intuitive way is to consider the pairs of values and then assign each of the 36 outcomes in the sample space a number. For our first example, let X be the "sum of the two values that land face up." This process allows us to assign a number to each outcome in our sample space, illustrated below.

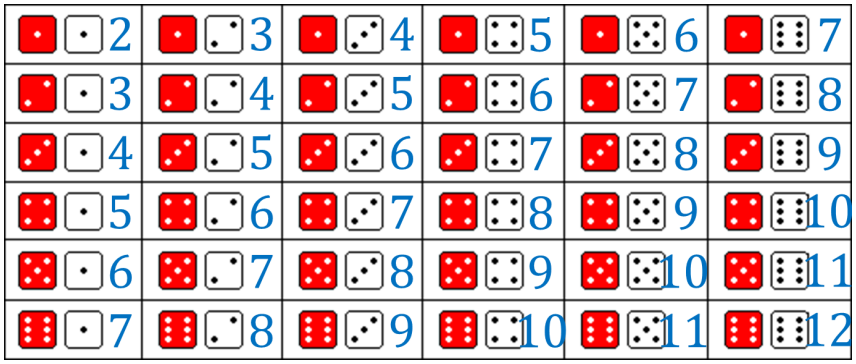


Figure 4.1.1: Sum of the two values that land face up when rolling two fair dice

As we can see in the figure above, the random variable X has 11 possible values: $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Our next task is to determine the probabilities for each value that X can be. For our example, we have already computed some of the probabilities; see Text Exercises 3.1.3.3and 3.3.2.1. By selecting a value of our random variable, say $X = 3$, we define a specific

event in the sample space of our random experiment (namely, rolling a 1 and 2) and rely on the content of the previous chapter. So $P(X=3) \quad = P(1 \text{ and } 2) \quad = P(1 \text{ FIRST and } 2 \text{ SECOND or } 2 \text{ FIRST and } 1 \text{ SECOND}) \quad = \frac{1}{36}+\frac{1}{36}-\frac{0}{36} \quad = \frac{2}{36} \quad = \frac{1}{18}$. We encourage the reader to confirm each of the probabilities in the table below.

Table 4.1.1: Probability distribution of the random variable X

$X = x_j$	$P(X = x_j)$
2	$\frac{1}{36} \approx 2.7778\%$
3	$\frac{2}{36} = \frac{1}{18} \approx 5.5556\%$
4	$\frac{3}{36} = \frac{1}{12} \approx 8.3333\%$
5	$\frac{4}{36} = \frac{1}{9} \approx 11.1111\%$
6	$\frac{5}{36} \approx 13.8889\%$
7	$\frac{6}{36} = \frac{1}{6} \approx 16.6667\%$
8	$\frac{5}{36} \approx 13.8889\%$
9	$\frac{4}{36} = \frac{1}{9} \approx 11.1111\%$
10	$\frac{3}{36} = \frac{1}{12} \approx 8.3333\%$
11	$\frac{2}{36} = \frac{1}{18} \approx 5.5556\%$
12	$\frac{1}{36} \approx 2.7778\%$

? Text Exercise 4.1.1

When rolling a pair of fair dice, each die lands with a value face up. Construct the probability distribution for the random variable Y, defined to be "the maximum value of the two dice rolled."

Answer

Visualizing the sample space may help identify possible values and determine probabilities for some random variables. We will want to use deeper reasoning when our sample spaces are much larger. We use both visualization and reasoning in this example. To construct a probability distribution, we must first determine the possible values of our random variable Y and then determine their probabilities.

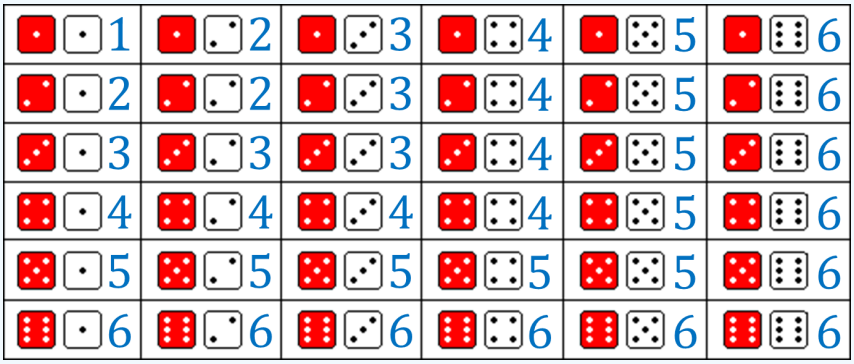


Figure 4.1.2 Maximum value of two fair dice rolled

From the figure above, we can tell that our random variable Y has 6 possible values: $\{1,2,3,4,5,6\}$. With 6 occurring most frequently. Since we are rolling fair dice, each outcome is equally likely, so we produce our probability distribution by counting the number of occurrences of each value. We make the probability in the table below.

Table 4.1.2 Probability distribution of the random variable Y

$Y = y_j$	$P(Y = y_j)$
1	$\frac{1}{36} \approx 2.7778\%$
2	$\frac{3}{36} = \frac{1}{12} \approx 8.3333\%$
3	$\frac{5}{36} \approx 13.8889\%$
4	$\frac{7}{36} \approx 19.4444\%$
5	$\frac{9}{36} = \frac{1}{4} = 25\%$
6	$\frac{11}{36} \approx 30.5556\%$

Let us reason our way to the probability distribution without listing all the outcomes. Knowing that the values on our dice range from 1 to 6, we can restrict our considerations for the possible values of Y to these; that is, Y has 6 possible values: $\{1,2,3,4,5,6\}$.

To determine $P(Y = 6)$, we note that 6 is the largest value on our dice. So if a 6 is rolled, it is the maximum. So $P(Y = 6) = P(6 \text{ IS ROLLED FIRST OR SECOND}) = \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}$.

To determine $P(Y = 5)$, we note that the only number larger than 5 is 6. So $P(Y = 5) = P(5 \text{ IS ROLLED and } 6 \text{ IS NOT ROLLED}) = \frac{11}{36} \cdot \frac{9}{11} = \frac{9}{36} = \frac{1}{4}$. Notice that $P(6 \text{ IS NOT ROLLED} | 5 \text{ IS ROLLED}) = \frac{9}{11}$ because there are 11 ways to roll a 5 and 9 of them do not contain a 6.

A similar process could continue through all of the possible values, but we might notice an easier way to count; for any particular value y_j of Y, we have the outcome of rolling double y_js, and the remaining outcomes come in pairs. The number of pairs equals the number of values less than y_j. We develop a formula for our probabilities: $P(Y = y_j) = \frac{2 \cdot (y_j - 1) + 1}{36}$. Check that our reasoning produces the same probability distribution as above.

Note: Probabilities Across all Possible Values

Recall that the sum of the probabilities of all outcomes from a sample space is 1. This is true because when conducting a random experiment, something must happen; a single outcome from the sample space must occur, and no two outcomes in the sample space are the same. If we have n outcomes, $1 = P(\text{OUTCOME}_1 \text{ or } \text{OUTCOME}_2 \text{ or } \ldots \text{ or } \text{OUTCOME}_n) = P(\text{OUTCOME}_1) + P(\text{OUTCOME}_2) + \ldots + P(\text{OUTCOME}_n)$.

A similar line of reasoning follows for random variables. Since a random variable assigns a number to every outcome, and an outcome must occur when a random experiment is conducted, we are sure that some value will occur and no outcome will return two values. If we have n values for a random variable X, $1 = P(X = x_1 \text{ or } X = x_2 \text{ or } \ldots \text{ or } X = x_n) = P(X = x_1) + P(X = x_2) + \ldots + P(X = x_n)$. The sum of probabilities across all possible values of a random variable must always equal 1. This means the sum of all the values in the $P(X = x)$ column of a probability distribution must add up to 1.

Let us confirm the sum of the probability column of a probability distribution is 1 for the random variables that we have discussed thus far, X and Y.

X: $\frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{36}{36} = 1$

Y: $\frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{7}{36} + \frac{9}{36} + \frac{11}{36} = \frac{36}{36} = 1$

Text Exercise 4.1.2

Consider the random variable D, loosely based on this article from 2009, which returns the number of days adults exercise in a week and its incomplete probability distribution below.

Table 4.1.3: Incomplete probability distribution for the random variable D

$D = d_j$	$P(D = d_j)$
0	0.28
1	0.11
2	
3	0.14
4	0.10
5	0.15
6	0.08
7	0.04

1. Complete the probability distribution by determining $P(D = 2)$.

Answer

The sum of all the probabilities in a probability distribution must equal 1. We can compute $P(D = 2)$ by figuring out what value makes the sum 1. $0.28+0.11+0.14+0.10+0.15+0.08+0.04 = .90$. So $P(D = 2) = 1 - 0.90 = 0.10$. Ten percent of adults exercise for just 2 days a week.

2. What is the probability that a randomly selected adult exercises at least 5 days a week?

Answer

We are trying to determine the probability that a randomly selected adult exercises 5, 6, or 7 days a week. which we can denote as $P(D \geq 5) = P(D = 5 \text{ or } D = 6 \text{ or } D = 7)$. Since each event is assigned a single value, they are mutually exclusive; thus, we can simply add the probabilities of each event. $P(D \geq 5) = P(D = 5 \text{ or } D = 6 \text{ or } D = 7) = P(D = 5) + P(D = 6) + P(D = 7) = 0.15 + .08 + .04 = 0.27$. We understand this to mean that 27% of adults work out at least 5 days a week.

3. Determine and explain the meaning of $P(2 < D \leq 4)$.

Answer

We are trying to determine the probability that a randomly selected adult exercises more than 2 times a week but no more than 4 times a week or equivalently that a randomly selected adult exercises 3 or 4 days a week. $P(2 < D \leq 4) = P(D = 3) + P(D = 4) = 0.14 + 0.10 = 0.24$. So 24% of adults exercise 3 or 4 days a week.

Now consider the random experiment of rolling two fair dice from a slightly different perspective and arrive at another type of random variable. We could define a random variable Z to be "the time (in seconds) it takes both dice to come to a complete stop after one die leaves our hands." We understand our random variable by examining the possible values that our random variable takes on. Determining the precise values is difficult. What is the shortest time? Does it always take at least 1 second? What is the longest time? Can it ever exceed 5 seconds? We cannot give definitive answers. However, after a moment or two of thought, we recognize that the possible values will take on any numerical value in an interval of positive real numbers. Hopefully, this last description reminds us of a type of variable. In Chapter 1, we defined two types of quantitative variables: discrete and continuous. Here, we make similar designations: **discrete random variable** and **continuous random variable**, based on the possible outcomes. Examples X, Y, and D from above are discrete random variables while Z is a continuous random variable. Our

understanding of probability needs further development to handle continuous random variables. This will take place in the latter portion of this chapter; for now, we restrict ourselves to the study of discrete random variables.

Text Exercise 4.1.3

Classify each random variable as either discrete or continuous. Explain.

1. P = the prescription count of a randomly chosen patient.

Answer

We understand a patient's prescription count to be the number of medicines prescribed. While a patient may be prescribed a half or double dosage, this is not half of a prescription. There are gaps between each possible value that P takes on, making P a discrete random variable.

2. V = the appraisal value of a randomly chosen coin collection.

Answer

An appraisal value must be given in some currency, perhaps U.S. dollars. Currencies have a smallest denomination. Therefore, there must be gaps between the possible values in the appraisal value, making V a discrete random variable.

3. T = the total distance traveled on the campaign trail of a randomly chosen politician.

Answer

A politician's total distance on the campaign trail (in any standard unit) may be any nonnegative number within a reasonable magnitude. T is a continuous random variable.

Discrete Uniform Distribution

As the name indicates, the probability distribution of a random variable explains how probabilities are distributed. We say a random variable has a **discrete uniform distribution** if the random variable is discrete and each outcome has equal probability. If we consider rolling a single fair die and define a random variable S to be the number that lands face up, the random variable S has a discrete uniform distribution. There are only six possible values making S discrete, and since the die is fair, each value is equally probable. $P(S=s) = \frac{1}{6}$ for any s in $\{1,2,3,4,5,6\}$.

Text Exercise 4.1.4

1. Consider the discrete random variable R, with 10 values, that has a discrete uniform distribution, and determine the probability of each value of R.

Answer

Since R has a discrete uniform distribution and takes on 10 values we have that $P(R=r_1) = P(R=r_2) = \ldots = P(R=r_{10})$ and $\sum_{j=1}^{10} P(R=r_j) = 1$. Combining these yields that $1 = \sum_{j=1}^{10} P(R=r_j) = \sum_{j=1}^{10} P(R=r_1) = 10 \cdot P(R=r_1)$ meaning $P(R=r_1) = \frac{1}{10}$ and thus $P(R=r_j) = \frac{1}{10}$ for any j in $\{1,2,3,\ldots,10\}$.

2. Consider the discrete random variable R which takes on k values and has a discrete uniform distribution, determine the probability of each value that R takes on.

Answer

Since R has a discrete uniform distribution and takes on k values we have that $P(R=r_1) = P(R=r_2) = \cdots = P(R=r_k)$ and $\sum_{j=1}^{k} P(R=r_j) = 1$. Combining these yields that $1 = \sum_{j=1}^{k} P(R=r_j) = \sum_{j=1}^{k} P(R=r_1) = k \cdot P(R=r_1)$ meaning $P(R=r_1) = \frac{1}{k}$ and thus $P(R=r_j) = \frac{1}{k}$ for any j in $\{1,2,3,\ldots,k\}$.

3. Consider the random variables X, Y, and D which have been recurring this section. For each of them, determine, with justification, if they are a uniform random variable or not.

Answer

First consider X. The probability that X is 7 is not the same as the probability that X is 12. Therefore, X is not a uniformly distributed random variable. We should be able to convince ourselves, using similar reasoning, that Y and D are also both not uniformly distributed.

4.2: Analyzing Discrete Random Variables

Learning Objectives

- Draw connections between a relative frequency distribution of a data set and a probability distribution of a random variable
- Define, compute, and interpret the expected value of a discrete random variable
- Define and compute the variance and standard deviation of a discrete random variable
- Interpret the standard deviation of a discrete random variable

Section 4.2 Excel File (contains all of the data sets for this section)

Random Variables and Data

As we seek to understand the world around us, there are times when we are unsure what could happen in a particular situation; we do not necessarily know all the possible outcomes or the likelihood of every outcome happening. That is okay. We learn by observation, data collection, and experimentation. When studying a quantitative variable, we collect data by measuring a value for each observation. Once our data is collected, we can analyze the data using relative frequency distributions to show the proportion of the observations in a particular class and provide an empirical estimate of the probability of that class occurring. We realize that our relative frequency distributions are connected with the probability distributions of our random variables (classes corresponding to values the variable takes on and relative frequencies corresponding to probabilities). There is, however, a significant difference to keep in mind. Relative frequency distributions describe data that has been collected. Random variables describe the possibilities and probabilities of what can happen. Relative frequency distributions are descriptive, while random variables are predictive.

To analyze and understand data, we developed methods of visualization, measures of centrality, and measures of dispersion. For these methods with random variables, we use the connection to relative frequency distributions. Recall how we treated these three concepts when given a relative frequency distribution for a discrete quantitative variable.

Visualization: We visualized the relative frequency distribution using bar graphs.

Measures of Centrality: We discovered that the mean of a data set could be calculated by "weighing" our class values (x_j) by their relative frequencies $(P(x_j))$

$$\mu = \frac{\sum [x_j \cdot P(x_j)]}{\sum P(x_j)} = \sum [x_j \cdot P(x_j)]$$

Measures of Dispersion: We discovered that the variance of a population data set could be calculated by "weighing" the squared deviations from the mean $\left((x_j - \mu)^2\right)$ by their relative frequencies $(P(x_j))$

$$\sigma^2 = \frac{\sum \left[(x_j - \mu)^2 \cdot P(x_j)\right]}{\sum P(x_j)} = \sum \left[(x_j - \mu)^2 \cdot P(x_j)\right]$$

From the variance, we can also compute the standard deviation of the data set: $\sigma = \sqrt{\sigma^2}$.

We will visualize our discrete random variables with bar graphs and develop measures of centrality and dispersion similar to the mean and variance of a population data set.

Analyzing Discrete Random Variables

The primary challenge in connecting our analyses of relative frequency distributions and random variables is interpretation. As mentioned above, relative frequency distributions are constructed from collected data and describe what happened; random variables are constructed from all the possible values resulting from a random experiment and the associated probabilities of these values occurring. Even in this distinction, the best predictor of future behavior is past behavior; we often use relative frequencies as estimates of probabilities (the empirical method of probability). So, how can we understand the measures of centrality and dispersion for a random variable? They would be estimates for the actual measures of centrality and dispersion if we were to repeatedly run the random experiment, collect data, and analyze it. Because of the Law of Large Numbers, we would expect the estimates and computed measures to converge if we conducted the random experiment repeatedly. With this in mind, we will study discrete random variables.

Graphical Visualization: Bar Graphs

The method of constructing bar graphs remains the same for discrete random variables. The values of the random variable are listed on the horizontal axis, and bars are formed with the height indicating the probability of that particular value occurring. The gaps between the bars indicate the discrete nature of the random variable. Here, we produce bar graphs for the random variables X, Y, and D from our last section (the sum of two fair dice rolled, the max of two fair dice rolled, and the number of days adults exercise in a week, respectively).

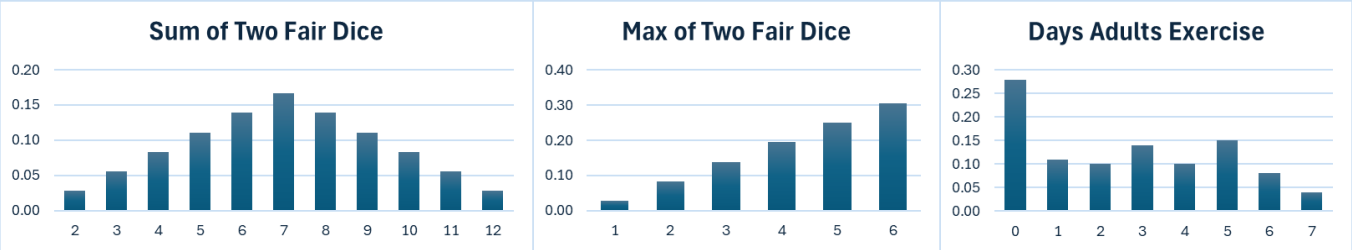


Figure 4.2.1: Bar graph representations of the probability distributions of the random variables X, Y, and D (left to right)

Measure of Centrality: Expected Value

Recall that the mean of a data set is the balancing point of that data set, which we could visualize from our graphs. Given this same manner of representation, the mean of a random variable X, also referred to as the **expected value of X, $\mathrm{E}(X)$,** is also the balance point of the bar graph. From the symmetry of X in the figure above, we conclude that $\mathrm{E}(X) = 7$. Determining the center of mass for Y and D is a little more complicated, but by looking at the distributions, we can see that $\mathrm{E}(Y) > 4$ and $\mathrm{E}(D) < 3$. A more precise calculation is in order. By understanding the relative frequency as an estimate of that particular event's probability, we define the **expected value of a discrete random variable X.**

$$\mu = \mathrm{E}(X) = \frac{\sum [x_j \cdot P(X = x_j)]}{\sum P(X = x_j)} = \sum [x_j \cdot P(X = x_j)]$$

In computing the expected value of a discrete random variable, we consider each possible value of the random variable and weigh it according to the associated probability; the higher the probability, the heavier the weight. Since the sum of probabilities of all possible values adds to 1, the denominator consisting of the "total" weight simplifies. Let us confirm that the expected value for the sum of two fair dice is indeed 7.

Table 4.2.1: Computation of the expected value of the sum of two fair dice

$X = x_j$	$P(X = x_j)$	$x_j \cdot P(X = x_j)$
2	$\frac{1}{36}$	$2 \cdot \frac{1}{36} = \frac{2}{36}$
3	$\frac{2}{36}$	$3 \cdot \frac{2}{36} = \frac{6}{36}$
4	$\frac{3}{36}$	$4 \cdot \frac{3}{36} = \frac{12}{36}$
5	$\frac{4}{36}$	$5 \cdot \frac{4}{36} = \frac{20}{36}$
6	$\frac{5}{36}$	$6 \cdot \frac{5}{36} = \frac{30}{36}$
7	$\frac{6}{36}$	$7 \cdot \frac{6}{36} = \frac{42}{36}$
8	$\frac{5}{36}$	$8 \cdot \frac{5}{36} = \frac{40}{36}$
9	$\frac{4}{36}$	$9 \cdot \frac{4}{36} = \frac{36}{36}$
10	$\frac{3}{36}$	$10 \cdot \frac{3}{36} = \frac{30}{36}$

$X = x_j$	$P(X = x_j)$	$x_j \cdot P(X = x_j)$
11	$\frac{2}{36}$	$11 \cdot \frac{2}{36} = \frac{22}{36}$
12	$\frac{1}{36}$	$12 \cdot \frac{1}{36} = \frac{12}{36}$
$\mu = \mathrm{E}(X) = \frac{2}{36} + \frac{6}{36} + \ldots + \frac{12}{36} = \frac{252}{36} = 7$		

After adding all of the entries ($x_j \cdot P(X = x_j)$) in the third column, we arrived at a total of 7, confirming our visual estimation. $\mathrm{E}(X) = 7$ means that as we roll a pair of fair dice repeatedly, we expect the mean of all of the sums to be close to 7. Consider rolling two fair dice using a simulator and calculating the mean of the dice sums. We (the authors) ran the simulation with 20 trials, producing a data set of 20 dice sums, and found the mean to be $\frac{142}{20} = 7.1$; this is fairly close to our expected value of 7. Run the simulation yourself and compute the mean. How close to 7 is your computed mean?

? Text Exercise 4.2.1

1. Recall the discrete random variable Y which describes the maximum value of two fair dice when rolled. Compute and interpret the expected value $\mathrm{E}(Y)$ using its probability distribution reproduced below.

Table 4.2.2: Probability distribution for the random variable Y

$Y = y_j$	$P(Y = y_j)$	$y_j \cdot (P(Y = y_j)$
1	$\frac{1}{36}$	
2	$\frac{3}{36}$	
3	$\frac{5}{36}$	
4	$\frac{7}{36}$	
5	$\frac{9}{36}$	
6	$\frac{11}{36}$	

Answer

Table 4.2.3 Table of computation

$Y = y_j$	$P(Y = y_j)$	$y_j \cdot (P(Y = y_j)$
1	$\frac{1}{36}$	$1 \cdot \frac{1}{36} = \frac{1}{36}$
2	$\frac{3}{36}$	$2 \cdot \frac{3}{36} = \frac{6}{36}$
3	$\frac{5}{36}$	$3 \cdot \frac{5}{36} = \frac{15}{36}$
4	$\frac{7}{36}$	$4 \cdot \frac{7}{36} = \frac{28}{36}$
5	$\frac{9}{36}$	$5 \cdot \frac{9}{36} = \frac{45}{36}$
6	$\frac{11}{36}$	$6 \cdot \frac{11}{36} = \frac{66}{36}$
$\mu = \mathrm{E}(Y) = \frac{1}{36} + \frac{6}{36} + \ldots + \frac{66}{36} = \frac{161}{36} = 4 + \frac{17}{36} \approx 4.4722$		

The expected value of the random variable Y is approximately 4.4722,meaning after repeatedly rolling two fair dice and examining the maximum values from each pair, the mean maximum value would be about 4.4722. We can verify this experimentally; see for yourself. Roll two dice, record the highest number, repeat 20 times, and then take the mean of those 20 numbers. We (the authors) obtained a mean of 4.7 when doing this. Notice that this seems less accurate than our empirical estimate obtained for X. Some random variables require more trials than others to accurately estimate the expected value. Note that our expected value cannot occur for any single trial of our random experiment.

2. Recall the discrete random variable D which describes the number of days adults exercise per week. Compute and interpret the expected value $\mathrm{E}(D)$ using its probability distribution reproduced below.

Table 4.2.4: Probability distribution of the random variable D

$\boldsymbol{D = d_j}$	$\boldsymbol{P(D = d_j)}$	$\boldsymbol{d_j \cdot (P(D = d_j)}$
0	0.28	
1	0.11	
2	0.10	
3	0.14	
4	0.10	
5	0.15	
6	0.08	
7	0.04	

Answer

Table 4.2.5: Table of computation

$D = d_j$	$P(D = d_j)$	$d_j \cdot (P(D = d_j)$
0	0.28	$0 \cdot 0.28 = 0$
1	0.11	$1 \cdot 0.11 = 0.11$
2	0.10	$2 \cdot 0.10 = 0.20$
3	0.14	$3 \cdot 0.14 = 0.42$
4	0.10	$4 \cdot 0.10 = 0.4$
5	0.15	$5 \cdot 0.15 = 0.75$
6	0.08	$6 \cdot 0.08 = 0.48$
7	0.04	$7 \cdot 0.04 = 0.28$
$\mu = \mathrm{E}(D) = 0 + 0.11 + \ldots + 0.28 = 2.64$		

The expected value of the random variable D is 2.64days, meaning that after repeatedly asking random adults how many days they work out per week, we would predict the mean number of days to be about 2.64.

Measures of Dispersion: Variance and Standard Deviation

A question arises from our discussion of expected value: how much variation should there be as the random experiment is repeated? Will most values be close to the expected value, or will a wide variety of values occur? We can answer these questions intuitively using bar graphs, but the intuition is rarely straightforward. Use Figure 4.2.1 to determine which random variable has the most spread before continuing with your reading.

Using the notion of relative frequency as an estimate for a value's probability and utilizing our expected value as our measure of center, we define the **variance of a discrete random variable X.**

$$\sigma^2 = \mathrm{Var}(X) = \frac{\sum\left[(x_j - \mathrm{E}(X))^2 \cdot P(X = x_j)\right]}{\sum P(X = x_j)} = \sum\left[(x_j - \mathrm{E}(X))^2 \cdot P(X = x_j)\right]$$

$$= \frac{\sum\left[(x_j - \mu)^2 \cdot P(X = x_j)\right]}{\sum P(X = x_j)} = \sum\left[(x_j - \mu)^2 \cdot P(X = x_j)\right]$$

In computing the variance of a discrete random variable, we "weigh" the values that we are averaging, namely the squared deviations from the mean, by their probabilities. Since the sum of the probabilities of all the possible values is 1, the denominator reduces. As before, we define the **standard deviation of a discrete random variable X** as the square root of the random variable's variance. Let us determine the dispersion of the discrete random variables, X, Y, and D by computing their variances and standard deviations.

Table 4.2.6: Table of computation

$X = x_j$	$P(X = x_j)$	$(x_j - \mu)^2$	$(x_j - \mu)^2 \cdot P(X = x_j)$
2	$\frac{1}{36}$	$(2-7)^2 = 25$	$25 \cdot \frac{1}{36} = \frac{25}{36}$
3	$\frac{2}{36}$	$(3-7)^2 = 16$	$16 \cdot \frac{2}{36} = \frac{32}{36}$
4	$\frac{3}{36}$	$(4-7)^2 = 9$	$9 \cdot \frac{3}{36} = \frac{27}{36}$
5	$\frac{4}{36}$	$(5-7)^2 = 4$	$4 \cdot \frac{4}{36} = \frac{16}{36}$
6	$\frac{5}{36}$	$(6-7)^2 = 1$	$1 \cdot \frac{5}{36} = \frac{5}{36}$
7	$\frac{6}{36}$	$(7-7)^2 = 0$	$0 \cdot \frac{6}{36} = \frac{0}{36}$
8	$\frac{5}{36}$	$(8-7)^2 = 1$	$1 \cdot \frac{5}{36} = \frac{5}{36}$
9	$\frac{4}{36}$	$(9-7)^2 = 4$	$4 \cdot \frac{4}{36} = \frac{16}{36}$
10	$\frac{3}{36}$	$(10-7)^2 = 9$	$9 \cdot \frac{3}{36} = \frac{27}{36}$
11	$\frac{2}{36}$	$(11-7)^2 = 16$	$16 \cdot \frac{2}{36} = \frac{32}{36}$
12	$\frac{1}{36}$	$(12-7)^2 = 25$	$25 \cdot \frac{1}{36} = \frac{25}{36}$
$\sigma^2 = \mathrm{Var}(X) = \frac{25}{36} + \frac{32}{36} + \ldots + \frac{25}{36} = \frac{210}{36} = 5\frac{5}{6} \approx 5.8333$			
$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{35}{6}} \approx 2.4152$			

As with the mean, we understand the standard deviation as an estimate for the computed standard deviation as we repeatedly conduct the random experiment. As repetitions increase, the computed standard deviation approaches the expected standard deviation value. We understand the mean of a data set as its center and can loosely understand the standard deviation as the typical distance our observations are from the mean. If we translated this in terms of random variables, we could say that a typical value is within a standard deviation from the expected value. Within the context of summing the rolls of two fair dice, we would expect a typical value to be within 2.4152units from 7, those sums being 5, 6, 7, 8, or 9. If we were playing a board game based on the sum of the rolls of two fair dice, we could expect these 5 numbers to occur somewhat "regularly" and plan our strategy accordingly.

Text Exercise 4.2.2

1. Determine the probability that the sum of the rolling of two fair dice would be within one standard deviation of the expected value: $P(|X-\mu|<\sigma)$.

Answer

Since the mean is 7 and the standard deviation is about 2.4152 we are looking for the probability that our random variable is between the values between $7-2.4152=4.5848$ and $7+2.4152=9.4152$. Our question reduces to finding $P(X=5,6,7,8,\text{ or } 9)=\frac{4}{36}+\frac{5}{36}+\frac{6}{36}+\frac{5}{36}+\frac{4}{36}=\frac{24}{36}=\frac{2}{3}\approx 66.6667\%$.

2. Determine the probability that the sum of the rolling of two fair dice would be within two standard deviations of the expected value: $P(|X-\mu|<2\sigma)$.

Answer

We are now looking for the probability that our random variable takes on the values between $7-2\cdot 2.4152=2.1696$ and $7+\cdot 2.4152=11.8304$. So our question reduces to finding $P(3\leq X\leq 11)=1-P(X=2 \text{ or } 12)=1-(\frac{1}{36}+\frac{1}{36})=1-\frac{2}{36}=\frac{34}{36}=\frac{17}{18}\approx 94.4444\%$.Notice that all possible values of our random variable lie within 3 standard deviations of the expected value.

Text Exercise 4.2.3

1. Recall the discrete random variable Y that describes the maximum value of two fair dice when rolled. In text exercise 4.2.1.1, we computed that $\mathrm{E}(Y)=\frac{161}{36}$.
   a. Compute the variance and standard deviation of Y using its probability distribution reproduced below.
   b. Compute $P(|Y-\mu|<\sigma)$.
   c. Compute $P(|Y-\mu|<2\sigma)$.

Table 4.2.7: Probability distribution of the random variable Y

$Y=y_j$	$P(Y=y_j)$	$(y_j-\mu)^2\cdot P(Y=y_j)$
1	$\frac{1}{36}$	
2	$\frac{3}{36}$	
3	$\frac{5}{36}$	
4	$\frac{7}{36}$	
5	$\frac{9}{36}$	
6	$\frac{11}{36}$	

Answer

Table 4.2.8 Table of computation

$Y=y_j$	$P(Y=y_j)$	$(y_j-\mu)^2\cdot P(Y=y_j)$
1	$\frac{1}{36}$	$\left(1-\frac{161}{36}\right)^2\cdot\frac{1}{36}\approx 0.3349$
2	$\frac{3}{36}$	$\left(2-\frac{161}{36}\right)^2\cdot\frac{3}{36}\approx 0.5093$

$Y=y_j$	$P(Y=y_j)$	$(y_j-\mu)^2\cdot P(Y=y_j)$
3	$\frac{5}{36}$	$\left(3-\frac{161}{36}\right)^2\cdot\frac{5}{36}\approx 0.3010$
4	$\frac{7}{36}$	$\left(4-\frac{161}{36}\right)^2\cdot\frac{7}{36}\approx 0.0434$
5	$\frac{9}{36}$	$\left(5-\frac{161}{36}\right)^2\cdot\frac{9}{36}\approx 0.0696$
6	$\frac{11}{36}$	$\left(6-\frac{161}{36}\right)^2\cdot\frac{11}{36}\approx 0.7132$
$\sigma^2=\text{Var}(Y)\approx 0.3349+0.5093+0.3010+0.0434+0.0696+0.7132\approx 1.9715$		
$\sigma=\sqrt{\text{Var}(Y)}\approx\sqrt{1.9715}\approx 1.4041$		

a. Note: when conducting such involved and messy calculations, be sure to round only at the very last stage. We provide intermediate approximations along the way to help facilitate readability, but we do not use the rounded values in the actual computations for final answers. The variance of Y is approximately $1.9715$and the standard deviation is about 1.4041

b. We are tasked with determining the probability that the value produced in running the random experiment will fall within one standard deviation of the expected value.

$$\begin{aligned} P\left(|Y-\mu|<\sigma\right) &= P\left(4.4722-1.4041<Y<4.4722+1.4041\right)=P\left(3.0681<Y<5.8763\right)\\ &= P\left(Y=4 \text{ or } 5\right)\\ &= \frac{7}{36}+\frac{9}{36}=\frac{16}{36}\\ &= \frac{4}{9}\approx 44.4444\% \end{aligned}$$

c. We now determine the probability that the value produced in running the random experiment will fall within two standard deviations of the expected value.

$$\begin{aligned} P\left(|Y-\mu|<2\sigma\right) &= P\left(4.4722-2\cdot 1.4041<Y<4.4722+2\cdot 1.4041\right)=P\left(1.664<Y<7.2804\right)\\ &= P\left(Y=2,3,4,5, \text{ or } 6\right)\\ &= 1-P\left(Y=1\right)=1-\frac{1}{36}\\ &= \frac{35}{36}\approx 97.2222\% \end{aligned}$$

2. Recall the discrete random variable D that describes the number of days adults exercise per week. In text exercise 4.2.1.2 we computed that $\text{E}(D)=2.64$ days.
   a. Compute the variance and standard deviation of D using its probability distribution reproduced below.
   b. Compute $P\left(|D-\mu|<\sigma\right)$.
   c. Compute $P\left(|D-\mu|<2\sigma\right)$.

Table 4.2.9: Probability distribution of the random variable D

$D=d_j$	$P(D=d_j)$	$(d_j-\mu)^2\cdot P(D=d_j)$
0	0.28	
1	0.11	
2	0.10	
3	0.14	
4	0.10	
5	0.15	

$D = d_j$	$P(D = d_j)$	$(d_j - \mu)^2 \cdot P(D = d_j)$
6	0.08	
7	0.04	

Answer

Table 4.2.10 Table of computation

$D = d_j$	$P(D = d_j)$	$(d_j - \mu)^2 \cdot P(D = d_j)$
0	0.28	$(0 - 2.64)^2 \cdot 0.28 \approx 1.9515$
1	0.11	$(1 - 2.64)^2 \cdot 0.11 \approx 0.2959$
2	0.10	$(2 - 2.64)^2 \cdot 0.10 \approx 0.0410$
3	0.14	$(3 - 2.64)^2 \cdot 0.14 \approx 0.0181$
4	0.10	$(4 - 2.64)^2 \cdot 0.10 \approx \backslash 0.1850$
5	0.15	$(5 - 2.64)^2 \cdot 0.15 \approx 0.8354$
6	0.08	$(6 - 2.64)^2 \cdot 0.08 \approx 0.9032$
7	0.04	$(7 - 2.64)^2 \cdot 0.04 \approx 0.7604$
$\sigma^2 = \text{Var}(D) \approx 1.9515 + 0.2959 + \cdots + 0.7604 \approx 4.9904$		
$\sigma = \sqrt{\text{Var}(D)} \approx \sqrt{4.9904} \approx 2.2339$		

a. The variance of D is approximately 4.9904days^2 and the standard deviation is about 2.2339days.
b. We determine the probability that the value produced in running the random experiment will fall within one standard deviation of the expected value.

$$\begin{aligned} P(|D - \mu| < \sigma) &= P(2.64 - 2.2339 < D < 2.64 - 2.2339) = P(0.4061 < D < 4.8739) \\ &= P(D = 1, 2, 3, \text{ or } 4) \\ &= 0.11 + 0.10 + 0.14 + 0.10 = 45\% \end{aligned}$$

c. We determine the probability that the value produced in running the random experiment will fall within two standard deviations of the expected value.

$$\begin{aligned} P(|D - \mu| < 2\sigma) &= P(2.64 - 2 \cdot 2.2339 < D < 2.64 - 2 \cdot 2.2339) = P(-1.8278 < D < 7.1078) \\ &= P(D = 0, 1, 2, 3, 4, 5, 6 \text{ or } 7) \\ &= 100\% \end{aligned}$$

As we have seen with the last several text exercises, the probability that a discrete random variable is within a standard deviation or two of the expected value varies depending on the distribution. This, hopefully, does not come as a surprise (if it did, recall our discussion in section 2.7 about Chebyshev's Inequality and the Empirical Rule). For this reason, we must take our understanding of a typical value rather loosely at this stage. It will get better. We end this section with text exercises exploring games of chance with cash prizes.

Note: Equivalent Formula for Variance

We may further reduce the formula for the variance of a random variable and produce an equivalent formulation with some computational benefits. The two formulas produce the same values; they are equivalent. The first formula emphasizes the underlying logic of what the measure means; we, therefore, recommend its use primarily, but we provide this second formula regardless.

$$\sigma^2 = \text{Var}(X) = \left(\sum \left[x_j^2 \cdot P(X = x_j)\right]\right) - \mu^2$$

Optional derivation for the mathematically inclined

$$\begin{aligned}\sigma^2 = \text{Var}(X) &= \sum \left[(x_j - \mu)^2 \cdot P(X = x_j)\right] \\ &= \sum \left[(x_j^2 - 2x_j\mu + \mu^2) \cdot P(X = x_j)\right] \\ &= \sum \left[x_j^2 \cdot P(X = x_j) - 2x_j\mu \cdot P(X = x_j) + \mu^2 \cdot P(X = x_j)\right] \\ &= \sum \left[x_j^2 \cdot P(X = x_j)\right] - \sum \left[2x_j\mu \cdot P(X = x_j)\right] + \sum \left[\mu^2 \cdot P(X = x_j)\right] \\ &= \sum \left[x_j^2 \cdot P(X = x_j)\right] - 2\mu \sum \left[x_j \cdot P(X = x_j)\right] + \mu^2 \sum P(X = x_j) \\ &= \sum \left[x_j^2 \cdot P(X = x_j)\right] - 2\mu^2 + \mu^2 \\ &= \left(\sum \left[x_j^2 \cdot P(X = x_j)\right]\right) - \mu^2\end{aligned}$$

Games of Chance

? Text Exercise 4.2.4

1. Consider entering a raffle with a single cash prize of $\$1,000$. Each player may only purchase one ticket for $\$25$, and only 100 raffle tickets will be sold.
   a. Represent our financial net gains or losses with the discrete random variable R and construct its probability distribution.
   b. Compute and interpret $\text{E}(R)$.

Answer

a. The random experiment is entering the raffle; either we win or we lose. If we lose, we are out the $\$25$. If we win, our net gain is $\$975$ because we spent $\$25$ on the raffle ticket. Our random variable R takes on two values: -25 and 975. Each ticket is typically equally likely to be drawn. Therefore, $P(X = -25)$ (the probability that we lose), is $\frac{99}{100}$. Its complement, the probability that we win, is $\frac{1}{100}$.

Table 4.2.11: Probability distribution for the random variable R

$R = r_j$	$P(R = r_j)$
-25	$\frac{99}{100}$
975	$\frac{1}{100}$

b. Table 4.2.12 Table of computation

$R = r_j$	$P(R = r_j)$	$r_j \cdot P(R = r_j)$
-25	$\frac{99}{100}$	$-25 \cdot \frac{99}{100} = -\frac{99}{4}$
975	$\frac{1}{100}$	$975 \cdot \frac{1}{100} = \frac{39}{4}$
$\mu = \text{E}(R) = -\frac{99}{4} + \frac{39}{4} = -\frac{60}{4} = -15$		

We have found that the expected value for this raffle is $-\$15$. When we play, we either lose $\$25$ or gain $\$975$, but on average we should expect to lose $\$15$. It is important to note that a negative expected value does not necessarily mean we will lose money. Someone will win the raffle. What it means is that, if we played this raffle many times, we would expect to lose money in the long run. Imagine if we did the raffle 100 times; we might expect to win 1 time and lose 99 times. In such a situation, we won $\$1000$, but we also lost $\$25 \cdot 100 = \$2500$ in purchasing tickets. Hence, we see a net loss of $\$1500$, for an average loss of $\$15$ each time.

2. Consider entering a raffle with one $\$1,000$ cash prize and one $\$500$ cash prize. Each player may only purchase one ticket for $\$50$, and 200 raffle tickets will be sold.
   a. Represent our financial net gains or losses with the discrete random variable R and construct its probability distribution.

b. Compute and interpret $\mathrm{E}(R)$.

Answer

a. We will set up our random variable similarly. However, this time, there are three possible values since there are three different outcomes: winning the big prize, winning the small prize, and losing altogether.

Table 4.2.13 Probability distribution for the random variable R

$R = r_j$	$P(R = r_j)$
-50	$\frac{198}{200}$
450	$\frac{1}{200}$
950	$\frac{1}{200}$

b. Table 4.2.14 Table of computation

$R = r_j$	$P(R = r_j)$	$r_j \cdot P(R = r_j)$
-50	$\frac{198}{200}$	$-50 \cdot \frac{198}{200} = -\frac{198}{4} =$
450	$\frac{1}{200}$	$450 \cdot \frac{1}{200} = \frac{9}{4}$
950	$\frac{1}{200}$	$950 \cdot \frac{1}{200} = \frac{19}{4}$
$\mu = \mathrm{E}(R) = -\frac{198}{4} + \frac{9}{4} + \frac{19}{4} = -\frac{170}{4} = -42.50$		

We have found that the expected value for this raffle is \-$42.50. Despite more prize money being available, the higher price of entering the raffle and the larger number of tickets sold made the expected value decrease starkly.

3. A family is hosting an extended family reunion and thought having a raffle with a single cash prize would be fun. The dad does not want to make or lose any money in running the raffle, but his wife would like the raffle to help cover the cost of hosting the reunion and wants $5 per ticket to help cover the costs. If 80 family members signed up for the raffle at a ticket price of $15, determine the cash prize and compute the expected value when

   a. the dad gets his desire.
   b. the mom gets her desire.

Answer

a. Since the dad does not want the raffle to make or lose any money on the raffle, the cash prize will need to be all the ticket revenue. Since there are 80 tickets being sold at $15 a ticket. The total ticket revenue is $1200. We can construct a random variable similar to the variables above representing the game from the player's perspective.

Table 4.2.15 Probability distribution for the random variable R along with computation

$R = r_j$	$P(R = r_j)$	$r_j \cdot P(R = r_j)$
-15	$\frac{79}{80}$	$-15 \cdot \frac{79}{80}$
1185	\(\dfrac{1}{80}\)	$1185 \cdot \frac{1}{80}$
$\mu = \mathrm{E}(R) = -\frac{15 \cdot 79}{80} + \frac{1185}{80} = \frac{-1185 + 1185}{80} = 0$		

If the dad gets his desire and all the ticket revenue is given as the cash prize, the expected value would $0.

b. If the mom gets her desire to help cover the cost of hosting the reunion, $10 of every ticket will go to the cash prize, and $5 will go towards the reunion. With 80 tickets sold, the cash prize would be $80 \cdot 10 = 800$ dollars.

Table 4.2.16 Probability distribution for the random variable R along with computation

$R = r_j$	$P(R = r_j)$	$r_j \cdot P(R = r_j)$
-15	$\frac{79}{80}$	$-15 \cdot \frac{79}{80}$
785	\(\dfrac{1}{80}\\)	$785 \cdot \frac{1}{80}$
$\mu = \mathrm{E}(R) = -\frac{15 \cdot 79}{80} + \frac{785}{80} = \frac{-1185 + 785}{80} = -\frac{400}{80} = -5$		

If the mom gets her desire, the expected value would $-\$5$. We notice that the dad did not want to make any money off the raffle and the expected value was $\$0$, while the mom wanted to make $\$5$ per ticket and the expected value was $-\$5$. In general, if the host wanted to make $\$x$ off of each ticket, then the expected value from the participant's perspective would be $-\$x$.

Having worked through these text exercises, we understand the expected value as a measure of who the game of chance favors: the players or the house/host. Under the mom's desires, the host family making $\$5$ a ticket is represented in the expected value being $-\$5$. If we then apply this idea to the earlier raffles, where the expected values were $-\$15$ and $-\$42.50$, we understand that the hosts were raising money at $\$15$ and $\$42.50$ a ticket. This is reasonable because raffles are generally run as fundraisers for some organizations.

? Text Exercise 4.2.5

Over the years, there have been big lottery prizes, as in millions and billions of dollars. Consider a simplified version of the Powerball game. While in reality, there are 9 different ways to win a cash prize from purchasing a $\$2$ ticket, we will only consider winning the grand prize without any additional options and assume there is only one winner. Recall that the grand prize is won when a player matches all 5 white balls (order does not matter) and the red Powerball. There are 69 white balls labeled 1 to 69 and 26 red balls labeled 1 to 26. The grand prize increases until a winner is found. Determine how large the grand prize must be for the expected value (from the player's perspective) to be nonnegative.

Answer

We are trying to find the smallest grand prize so that the expected value is nonnegative. Let us refer to the grand prize amount as p and set up a discrete random variable R modeling this simplified version of the Powerball. Refer to this previous text exercise for additional aid in determining the probabilities.

Table 4.2.17 Probability distribution for the random variable R along with computation

$R = r_j$	$P(R = r_j)$	$r_j \cdot P(R = r_j)$
-2	$1 - \frac{1}{292,201,338}$	$-2 + \frac{2}{292,201,338}$
$p - 2$	$\frac{1}{292,201,338}$	(\dfrac{p-2}{292,201,338}\)
$\mu = \mathrm{E}(R) = -2 + \frac{2}{292,201,338} + \frac{p}{292,201,338} - \frac{2}{292,201,338} = -2 + \frac{p}{292,201,338}$		

We want the expected value to be nonnegative, so we are looking to solve $\mathrm{E}(R) \geq 0$.

$$\begin{aligned} \mathrm{E}(R) &\geq 0 \\ -2 + \frac{p}{292,201,338} &\geq 0 \\ \frac{p}{292,201,338} &\geq 2 \\ p &\geq 2 \cdot 292,201,338 = 584,402,676 \end{aligned}$$

Only after the grand prize grows beyond $\$584,402,676$ will the expected value of the Powerball be positive. Note that this analysis does not consider taxation on the winnings, which is quite hefty.

Given the fact that expected value is negative (unless the pot is enormous), we know that playing the lottery is not a good way to make money.

Text Exercise 4.2.6

1. Consider a biased die, weighted so that rolling a one is 5 times as likely as the other sides, but each of the other sides are equally probable. Determine the probability distribution for the discrete random variable Z defined to be the value landing up after rolling this biased die.

Answer

We know that $P(Z=2) = P(Z=3) = P(Z=4) = P(Z=5) = P(Z=6)$, $P(Z=1) = 5P(Z=2)$, and that the sum of all the probabilities needs to be 1. We have $1 = P(Z=1) + P(Z=2) + P(Z=3) + P(Z=4) + P(Z=5) + P(Z=6) = 5P(Z=2) + P(Z=2) + P(Z=2) + P(Z=2) + P(Z=2) + P(Z=2) = 10P(Z=2)$. Thus $P(Z=2) = \frac{1}{10}$ and $P(Z=1) = \frac{5}{10} = \frac{1}{2}$.

Table 4.2.18 Probability distribution for the random variable Z

$Z = z_j$	$P(Z = z_j)$
1	$\frac{1}{2}$
2	$\frac{1}{10}$
3	$\frac{1}{10}$
4	$\frac{1}{10}$
5	$\frac{1}{10}$
6	$\frac{1}{10}$

2. Compare the expected value and standard deviation of Z with those of rolling a fair die.

Answer

Table 4.2.19 Table of computation for the two random variables

$Z = z_j$	$P(Z = z_j)$	$z_j \cdot P(Z = z_j)$	$(z_j - \mu)^2 \cdot P(Z = z_j)$	$F = f_j$	$P(F = f_j)$	$f_j \cdot P(F = f_j)$	$(f_j - \mu)^2 \cdot P(F =$
1	$\frac{1}{2}$	$\frac{1}{2}$	$\left(1-\frac{5}{2}\right)^2 \cdot \frac{1}{2} = \frac{9}{8}$	1	$\frac{1}{6}$	$\frac{1}{6}$	$\left(1-\frac{7}{2}\right)^2 \cdot \frac{1}{6} =$
2	$\frac{1}{10}$	$\frac{1}{5}$	$\left(2-\frac{5}{2}\right)^2 \cdot \frac{1}{10} = \frac{1}{40}$	2	$\frac{1}{6}$	$\frac{1}{3}$	$\left(2-\frac{7}{2}\right)^2 \cdot \frac{1}{6} =$
3	$\frac{1}{10}$	$\frac{3}{10}$	$\left(3-\frac{5}{2}\right)^2 \cdot \frac{1}{10} = \frac{1}{40}$	3	$\frac{1}{6}$	$\frac{1}{2}$	$\left(3-\frac{7}{2}\right)^2 \cdot \frac{1}{6} =$
4	$\frac{1}{10}$	$\frac{2}{5}$	$\left(4-\frac{5}{2}\right)^2 \cdot \frac{1}{10} = \frac{9}{40}$	4	$\frac{1}{6}$	$\frac{2}{3}$	$\left(4-\frac{7}{2}\right)^2 \cdot \frac{1}{6} =$
5	$\frac{1}{10}$	$\frac{1}{2}$	$\left(5-\frac{5}{2}\right)^2 \cdot \frac{1}{10} = \frac{5}{8}$	5	$\frac{1}{6}$	$\frac{5}{6}$	$\left(5-\frac{7}{2}\right)^2 \cdot \frac{1}{6} =$
6	$\frac{1}{10}$	$\frac{3}{5}$	$\left(6-\frac{5}{2}\right)^2 \cdot \frac{1}{10} = \frac{49}{40}$	6	$\frac{1}{6}$	1	$\left(6-\frac{7}{2}\right)^2 \cdot \frac{1}{6} =$

$Z = z_j$	$P(Z = z_j)$	$z_j \cdot P(Z = z_j)$	$(z_j - \mu)^2 \cdot P(Z = z_j)$	$F = f_j$	$P(F = f_j)$	$f_j \cdot P(F = f_j)$	$(f_j - \mu)^2 \cdot P(F =$
$\mu = \mathrm{E}(Z) = \frac{1}{2} + \frac{1}{5} + \ldots + \frac{3}{5} = \frac{25}{10} = \frac{5}{2} = 2.5$				$\mu = \mathrm{E}(F) = \frac{1}{6} + \frac{1}{3} + \ldots + 1 = \frac{21}{6} = \frac{7}{2} = 3.5$			
$\sigma^2 = \mathrm{Var}(Z) = \frac{9}{8} + \frac{1}{40} + \cdots + \frac{49}{40} = \frac{130}{40} = \frac{13}{4} = 3.25$				$\sigma^2 = \mathrm{Var}(F) = \frac{25}{24} + \frac{3}{8} + \cdots + \frac{25}{24} = \frac{70}{24} = \frac{35}{12} \approx 2.9167$			
$\sigma = \sqrt{3.25} \approx 1.8028$				$\sigma = \sqrt{\frac{35}{12}} \approx 1.7078$			

In comparing the two random variables, we notice that the bias brought the expected value closer to the biased value. This is relatively intuitive because one would expect the biased value to occur more often. The bias also increased the standard deviation a little. This may be less intuitive because it seems natural that the spread would decrease. After all, we expect ones to appear quite frequently. With a heavy enough bias, this can happen. The degree to which the expected value is pulled towards the biased value and the bias's impact on the standard deviation depends on the amount of bias and the probabilities of the other values.

4.3: Binomial Distributions

Learning Objectives

- Define the binomial random variable
- Construct binomial distributions
- Develop and use the probability distribution function for binomial random variables
- Provide and use alternative formulas for expected value and variance of binomial random variables
- Assess the necessity of independent trials

Binomial Random Variable

Suppose we had a weighted coin which only came up heads one-sixth of the time. If we flipped this coin 10 times, what is the probability that exactly 2 of those flips would be heads? In order to answer this question, we would need to recognize that each of the coin flips are independent and that the probability of getting heads is the same each time. The answer turns out to be about 29%; we will demonstrate how to compute this soon.

Now suppose we were rolling a fair die 10 times. What is the probability that exactly 2 of those rolls would be six? The astute reader may notice that the answer is the same: 29%. Why is that? Well, we either roll a six or we don't. Rolling a six is analogous to flipping heads and not rolling a six is analogous to flipping tails. Since each trial is independent and the probability of obtaining the outcome of interest is $\frac{1}{6}$, the two scenarios are the same from the perspective of probability.

In fact, we can be much more general. Suppose in a population consisting of millions of people, exactly one in six of them support some political candidate. If we randomly and independently selected 10 people, what is the probability that 2 of them would be supporters of the candidate? Suppose a factory produces light bulbs and one-sixth of them are dysfunctional. If an inspector were to randomly and independently select 10 of them, what is the probability that exactly 2 of them would be dysfunctional? Suppose an archer hits the bulls-eye with probability $\frac{1}{6}$ every time she shoots. If she takes 10 shots, what is the probability that she gets exactly 2 bulls-eyes? The answers to all of these questions are the same: 29%. It is clear that some of the details of the situations are irrelevant; all that matters is that a trial is repeated 10 times, each trial is independent, and the probability of the outcome of interest occurring is $\frac{1}{6}$ each time. These sorts of situations are the object of our discussion: binomial random variables.

Binomial distributions are the probability distributions for a particular type of discrete random variable: the binomial random variable. With binomial random variables, we are considering a single random experiment repeated, identically and independently, a fixed number of times. We call each repetition a trial and indicate the number of trials with n. As the adjective "binomial" indicates, we group the outcomes into two categories: successes and failures. The probability of a success on any given trial is denoted p, while the probability of a failure on any given trial is denoted q. Note that since we have only two categories covering the entire sample space, $q = 1 - p$. We define the **binomial random variable X** as the number of successes throughout all n trials. Every trial may fail, in which case, $X = 0$. On the other hand, every trial may be a success, in which case, $X = n$. In most cases, some trials will succeed while others fail. As such, X takes on any number in the set $\{0,1,2,3,\ldots,n\}$. In the examples discussed above, $n = 10$, $p = \frac{1}{6}$, $q = \frac{5}{6}$, and we were asking what is $P(X = 2)$.

Consider an example to help solidify these ideas. In a previous text exercise, we considered tossing a fair die three times and determined the probability of getting one or more throws landing with one face up. We can understand this situation as a binomial random variable. Our underlying random experiment is rolling a fair die. We fix the number of trials to 3. The trials are identical because we are similarly rolling the same die each time. The trials are independent because the outcomes of previous rolls do not affect current or future rolls. Since we are interested in rolling ones, we define that as a success. Rolling any other value (2, 3, 4, 5, or 6) constitutes a failure. We can easily compute the probabilities of success and failure on any individual trial; $p = \frac{1}{6}$ and $q = \frac{5}{6}$. We define our binomial random variable X to be the number of ones rolled in 3 tosses of a fair die. The possible values for X are 0, 1, 2, and 3. Recall that our interest in random variables lies in their probability distributions. We will now address constructing a binomial random variable's probability distribution.

Probability Distributions of Binomial Random Variables

We first build our intuition by constructing the probability distribution of our binomial random variable X : the number of ones rolled in 3 tosses of a fair die. When determining the probability for a particular value of a random variable, we generally considered all of the outcomes in the sample space and proceeded from there. Considering all three trials based on the values landing up would result in $6^3 = 216$ different outcomes. We can simplify our analysis by considering all three trials based on successes and failures; that is, rolling a one is considered a success and rolling anything other than a one is considered a failure. In this case, we only have $2^3 = 8$ considerations. We shall use S to indicate a trial with success and F to indicate a trial with a failure, and represent the possibility of a successful trial followed by failures on the second and third trials as SFF. This procedure is illustrated for several, but not all, possible outcomes in the figure below.

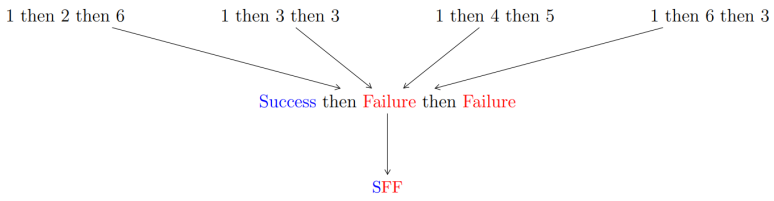


Figure 4.3.1: Outcomes of three rolls in succession understood in terms of successes and failures

The following figure groups the 8 possibilities by value of X and helps us build the probability distribution.

$X=0$	FFF
$X=1$	SFF FSF FFS
$X=2$	SSF SFS FSS
$X=3$	SSS

Table 4.3.1: Initial probability distribution for the random variable X

$X=x_j$	$P(X=x_j)$
0	$P(\text{FFF})$
1	$P(\text{SFF or FSF or FFS})$
2	$P(\text{SSF or SFS or FSS})$
3	$P(\text{SSS})$

Each trial occurs in sequence and is identical and independent; we can use both our addition and multiplication rules for probabilities to determine our probabilities. Remember that $P(\text{S})=p=\frac{1}{6}$ and $P(\text{F})=q=\frac{5}{6}$.

Table 4.3.2: Probability distribution for the random variable X

$X=x_j$	$P(X=x_j)$	
1	$\begin{aligned}P(\text{SFF or FSF or FFS}) &= P(\text{SFF})+P(\text{FSF})+P(\text{FFS})\\ &= P(\text{S})\cdot P(\text{F})\cdot P(\text{F})+P(\text{F})\cdot P(\text{S})\cdot P(\text{F})+P(\text{F})\cdot P(\text{F})\cdot P(\text{S})\\ &= 3P(\text{S})\cdot P(\text{F})^2\\ &= 3pq^2\end{aligned}$	$3\cdot\frac{1}{6}\cdot\left(\frac{5}{6}\right)^2=\frac{75}{216}\approx 34.7222\%$
3	$\begin{aligned}P(\text{SSS}) &= P(\text{S})\cdot P(\text{S})\cdot P(\text{S})\\ &= P(\text{S})^3\\ &= p^3\end{aligned}$	$\left(\frac{1}{6}\right)^3=\frac{1}{216}\approx 0.4630\%$
2	$\begin{aligned}P(\text{SSF or SFS or FSS}) &= P(\text{SSF})+P(\text{SFS})+P(\text{FSS})\\ &= P(\text{S})\cdot P(\text{S})\cdot P(\text{F})+P(\text{S})\cdot P(\text{F})\cdot P(\text{S})+P(\text{F})\cdot P(\text{S})\cdot P(\text{S})\\ &= 3P(\text{S})^2\cdot P(\text{F})\\ &= 3p^2q\end{aligned}$	$3\cdot\left(\frac{1}{6}\right)^2\cdot\frac{5}{6}=\frac{15}{216}\approx 6.9444\%$
0	$\begin{aligned}P(\text{FFF}) &= P(\text{F})\cdot P(\text{F})\cdot P(\text{F})\\ &= P(\text{F})^3\\ &= q^3\end{aligned}$	$\left(\frac{5}{6}\right)^3=\frac{125}{216}\approx 57.8704\%$

? Text Exercise 4.3.1

Consider the random variable Y that counts the number of heads in 4 flips of a fair coin. Verify that the random variable Y is a binomial random variable and construct its probability distribution.

Answer

The underlying random experiment is the flipping of the fair coin, which is to be repeated a fixed number of times; $n=4$. We are flipping the same coin in a similar fashion, meaning our trials are identical. We have independent trials because the outcome of one trial does not affect any of the other trials. We are counting the number of heads; in a successful trial, heads land face up, and alternatively, landing a tail would be a failure. Since we are using a fair coin, we have $P(\text{S})=P(\text{F})=\frac{1}{2}$ or $p=q=\frac{1}{2}$ confirming that Y is a binomial random variable. Since we have 4 trials, the set of possible values for Y is $\{0,1,2,3,4\}$. To construct the probability distribution, we consider the possible outcomes of all 4 trials in terms of successes and failures.

$Y=0$	FFFF
$Y=1$	SFFF FSFF FFSF FFFS
$Y=2$	SSFF SFSF SFFS FSSF FSFS FFSS
$Y=3$	SSSF SSFS SFSS FSSS
$Y=4$	SSSS

Table 4.3.3 Initial probability distribution for the random variable Y

$Y=y_j$	$P(Y=y_j)$
0	$P(\text{FFFF})$
1	$P(\text{SFFF or FSFF or FFSF or FFFS})$
2	$P(\text{SSFF or SFSF or SFFS or FSSF or FSFS or FFSS})$

$Y = y_j$	$P(Y = y_j)$
3	$P(\text{SSSF or SSFS or SFSS or FSSS})$
4	$P(\text{SSSS})$

Again, each trial occurs in sequence and is identical and independent; we use both our addition and multiplication rule for probabilities to determine our probabilities.

Table 4.3.4 Probability distribution for the random variable Y

$Y = y_j$	$P(Y = y_j)$	
0	$P(\text{FFFF}) = P(\text{F}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot$ $= P(\text{F})^4$ $= q^4$	$\left(\frac{1}{2}\right)^4 = \frac{1}{16} = 6.25\%$
1	$P(\text{SFFF or FSFF or FFSF or FFFS}) = P(\text{SFFF}) + P(\text{FSFF}) + P(\text{FFSF}) + P(\text{FFFS})$ $= P(\text{S}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot P(\text{F}) + P(\text{F}) \cdot P(\text{S}) \cdot P(\text{F}) \cdot P(\text{F})$ $+ P(\text{F}) \cdot P(\text{F}) \cdot P(\text{S}) \cdot P(\text{F}) + P(\text{F}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot P(\text{S})$ $= 4P(\text{S}) \cdot P(\text{F})^3$ $= 4pq^3$	$4 \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right)^3 = \frac{4}{16} = 25\%$
2	$P(\text{SSFF or SFSF or SFFS or FSSF or FSFS or FFSS}) = P(\text{SSFF}) + P(\text{SFSF})$ $+ P(\text{SFFS}) + P(\text{FSSF})$ $+ P(\text{FSFS}) + P(\text{FFSS})$ $= P(\text{S}) \cdot P(\text{S}) \cdot P(\text{F}) \cdot P(\text{F}) + P(\text{S}) \cdot P(\text{F}) \cdot P(\text{S}) \cdot P(\text{F})$ $+ P(\text{S}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot P(\text{S}) + P(\text{F}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot P(\text{F})$ $+ P(\text{F}) \cdot P(\text{S}) \cdot P(\text{F}) \cdot P(\text{S}) + P(\text{F}) \cdot P(\text{F}) \cdot P(\text{S}) \cdot P(\text{S})$ $= 6P(\text{S})^2 \cdot P(\text{F})^2$ $= 6p^2q^2$	$6 \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^2 = \frac{6}{16} = 37.5\%$
3	$P(\text{SSSF or SSFS or SFSS or FSSS}) = P(\text{SSSF}) + P(\text{SSFS}) + P(\text{SFSS}) + P(\text{FSSS})$ $= P(\text{S}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot P(\text{F}) + P(\text{S}) \cdot P(\text{S}) \cdot P(\text{F}) \cdot P(\text{S})$ $+ P(\text{S}) \cdot P(\text{F}) \cdot P(\text{S}) \cdot P(\text{S}) + P(\text{F}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot P(\text{S})$ $= 4P(\text{S})^3 \cdot P(\text{F})$ $= 4p^3q$	$4 \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right)^3 = \frac{4}{16} = 25\%$
4	$P(\text{SSSS}) = P(\text{S}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot$ $= P(\text{S})^4$ $= p^4$	$\left(\frac{1}{2}\right)^4 = \frac{1}{16} = 6.25\%$

Hopefully, we have noticed some patterns after building probability distributions for two binomial random variables. Let us formulate the patterns in the context of a general binomial random variable X with n trials and probability of success on any individual trial p. Recall that q, the probability of failure, is $1 - p$.

When we consider the possible outcomes of all trials in terms of successes and failures, the probabilities depend on the number of successes and failures, not on the order in which those successes and failures appear. Each event in $X = x_j$ has the same probability. If there are x_j successes, meaning we have $n - x_j$ failures, the probability of each event in $X = x_j$ is $p^{x_j}q^{n-x_j}$. All that is left to do is count the number of such events for any particular value x_j.

To count the number of ways that x_j successes can be assigned to the n trials, we can use combinations: ${}_nC_{x_j}$. We have ${}_nC_{x_j}$ many events in $X = x_j$ each with a probability of $p^{x_j}q^{n-x_j}$. Putting this all together, we arrive at a function that returns the probability of our binomial random variable. We call this the **probability distribution function for a binomial random variable X.**

$$P(X = x_j) = {}_nC_{x_j}\, p^{x_j} q^{n-x_j}$$

Check that the formula works by using it on the preceding example.

? Text Exercise 4.3.2

1. A virtual education company produces short multiple-choice quizzes for each content module. They currently have 5 questions with 2 options for each question. One school that uses this product worries about students passing these quizzes without learning the content. Determine the probability of a student obtaining an A or a B (obtaining at least an 80%,) on such a quiz by literally randomly guessing on each question.

Answer

We can understand this situation as a binomial random variable. We have a random experiment of a student randomly guessing on a multiple-choice question. The experiment is repeated 5 times because there are 5 questions. Since the student is randomly guessing on each question, the trials are identical and independent, with a probability of success at 50%. Let X be the binomial random variable that counts the number of correct guesses on these 5 question multiple-choice quizzes. This means a student needs 4 or 5 correct answers to obtain an A or a B. We need to find $P(X = 4 \text{ or } 5$. We can use the probability distribution function to find the answer. Recall that $n = 5$ and $p = 0.5$. Thus,

$$P(X=4) = {}_5C_4 \cdot \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right)^{5-4} = 5 \cdot \left(\frac{1}{16}\right) \cdot \left(\frac{1}{2}\right) = \frac{5}{32}$$

$$P(X=5) = {}_5C_5 \cdot \left(\frac{1}{2}\right)^5 \cdot \left(\frac{1}{2}\right)^{5-5} = 1 \cdot \left(\frac{1}{32}\right) \cdot \left(\frac{1}{2}\right)^0 = \frac{1}{32}$$

Thus, $P(X = 4 \text{ or } 5) = \frac{5}{32} + \frac{1}{32} = \frac{6}{32} = 18.75\%$.

2. Given the analysis in the first part of this text exercise, the virtual education company has decided to increase the number of options on each question while keeping the number of questions fixed at 5. They are considering using 3 or 4 options. Determine the probability that a student randomly guessing on a quiz will obtain an A or a B under both options.

Answer

Changing the number of options does not change the number of questions necessary, but it does change the probability of success on any given question. Let Y be the binomial random variable counting the number of correct guesses when there are 3 options on each question and Z be for 4 options. Thus, we are interested in $P(Y = 4 \text{ or } 5)$ and $P(Z = 4 \text{ or } 5)$. When there are 3 options, the probability of success, p_Y is only $\frac{1}{3}$. Similarly, when there are 4 options, the probability of success, p_Z is only $\frac{1}{4}$.

$$P(Y=4) = {}_5C_4 \cdot \left(\frac{1}{3}\right)^4 \cdot \left(\frac{2}{3}\right)^{5-4} = 5 \cdot \left(\frac{1}{81}\right) \cdot \left(\frac{2}{3}\right) = \frac{10}{243}$$

$$P(Y=5) = {}_5C_5 \cdot \left(\frac{1}{3}\right)^5 \cdot \left(\frac{2}{3}\right)^{5-5} = 1 \cdot \left(\frac{1}{243}\right) \cdot \left(\frac{2}{3}\right)^0 = \frac{1}{243}$$

$$P(Z=4) = {}_5C_4 \cdot \left(\frac{1}{4}\right)^4 \cdot \left(\frac{3}{4}\right)^{5-4} = 5 \cdot \left(\frac{1}{256}\right) \cdot \left(\frac{3}{4}\right) = \frac{15}{1024}$$

$$P(Z=5) = {}_5C_5 \cdot \left(\frac{1}{4}\right)^5 \cdot \left(\frac{3}{4}\right)^{5-5} = 1 \cdot \left(\frac{1}{1024}\right) \cdot \left(\frac{3}{4}\right)^0 = \frac{1}{1024}$$

Thus, $P(Y = 4 \text{ or } 5) = \frac{10}{243} + \frac{1}{243} = \frac{11}{243} \approx 4.5267\%$ and $P(Z = 4 \text{ or } 5) = \frac{15}{1024} + \frac{1}{1024} = \frac{16}{1024} \approx 1.5625\%$. Increasing the number of options significantly reduces the chances of a student obtaining an A or a B on a quiz by randomly selecting answers. We go from nearly 19% to just below 5% to just over 1%.

Expected Value, Variance, and Standard Deviation of Binomial Random Variables

Remember that binomial random variables are just a particular type of discrete random variable. That means everything we know about discrete random variables applies to binomial random variables. Binomial random variables have some very nice properties that make the calculations of expected value and variance much easier. Note that the formulas we develop here in this section only apply to binomial random variables and not all discrete random variables.

? Text Exercise 4.3.3

Using the definitions of expected value, variance, and standard deviation provided in the section on discrete random variables, determine these measures of centrality and dispersion for the binomial random variables: X being the number of ones rolled in 3 tosses of a fair die and Y being the number of heads in 4 flips of a fair coin.

Answer

These are the same random variables that we have been using throughout this section. We can utilize the probability distributions that we have already created.

Table 4.3.5 Table of computation for the random variable X

$X = x_j$	$P(X = x_j)$	$x_j \cdot P(X = x_j)$	$(x_j - \mu)^2 \cdot P(X = x_j)$
0	$\frac{125}{216}$	$0 \cdot \frac{125}{216} = 0$	$\left(0 - \frac{1}{2}\right)^2 \cdot \frac{125}{216} = \frac{1}{4} \cdot \frac{125}{216} = \frac{125}{864}$
1	$\frac{75}{216}$	$1 \cdot \frac{75}{216} = \frac{75}{216}$	$\left(1 - \frac{1}{2}\right)^2 \cdot \frac{75}{216} = \frac{1}{4} \cdot \frac{75}{216} = \frac{75}{864}$
2	$\frac{15}{216}$	$2 \cdot \frac{15}{216} = \frac{30}{216}$	$\left(2 - \frac{1}{2}\right)^2 \cdot \frac{15}{216} = \frac{9}{4} \cdot \frac{15}{216} = \frac{135}{864}$

$X = x_j$	$P(X = x_j)$	$x_j \cdot P(X = x_j)$	$(x_j - \mu)^2 \cdot P(X = x_j)$
3	$\frac{1}{216}$	$3 \cdot \frac{1}{216} = \frac{3}{216}$	$\left(3 - \frac{1}{2}\right)^2 \cdot \frac{1}{216} = \frac{25}{4} \cdot \frac{1}{216} = \frac{25}{864}$
$\mu = \mathrm{E}(X) = 0 + \frac{75}{216} + \frac{30}{216} + \frac{3}{216} = \frac{108}{216} = \frac{1}{2}$			
$\sigma^2 = \mathrm{Var}(X) = \frac{125}{864} + \frac{75}{864} + \frac{135}{864} + \frac{25}{864} = \frac{360}{864} = \frac{5}{12} \approx 0.4167$			
$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{5}{12}} \approx 0.6455$			

Table 4.3.6 Table of computation for the random variable Y

$Y = y_j$	$P(Y = y_j)$	$y_j \cdot P(Y = y_j)$	$(y_j - \mu)^2 \cdot P(Y = y_j)$
0	$\frac{1}{16}$	$0 \cdot \frac{1}{16} = 0$	$(0-2)^2 \cdot \frac{1}{16} = 4 \cdot \frac{1}{16} = \frac{1}{4}$
1	$\frac{1}{4}$	$1 \cdot \frac{1}{4} = \frac{1}{4}$	$(1-2)^2 \cdot \frac{1}{4} = 1 \cdot \frac{1}{4} = \frac{1}{4}$
2	$\frac{3}{8}$	$2 \cdot \frac{3}{8} = \frac{3}{4}$	$(2-2)^2 \cdot \frac{3}{8} = 0 \cdot \frac{3}{8} = 0$
3	$\frac{1}{4}$	$3 \cdot \frac{1}{4} = \frac{3}{4}$	$(3-2)^2 \cdot \frac{1}{4} = 1 \cdot \frac{1}{4} = \frac{1}{4}$
4	$\frac{1}{16}$	$4 \cdot \frac{1}{16} = \frac{1}{4}$	$(4-2)^2 \cdot \frac{1}{16} = 4 \cdot \frac{1}{16} = \frac{1}{4}$
$\mu = \mathrm{E}(Y) = 0 + \frac{1}{4} + \cdots + \frac{1}{4} = \frac{8}{4} = 2$			
$\sigma^2 = \mathrm{Var}(X) = \frac{1}{4} + \frac{1}{4} + \ldots + \frac{1}{4} = \frac{4}{4} = 1$			
$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{1} = 1$			

Having computed the expected value, variance, and standard deviation for two binomial random variables using the definitions, we now present quicker and easier methods for computing the expected value and variance. Just as with the alternative formula for the variance of a discrete random variable, these formulas are derived from our original definitions through mathematical simplification and produce the same values as the original definitions. We will not provide the work for this mathematical simplification but will provide a little intuition before providing the formulas. For example, if $p = 0.5$ and $n = 10$, then we are repeating a trial 10 times with probability of success being 0.5. We should expect, then, that half of the time, we will succeed. This means $E(X) = 0.5 \cdot 10 = 5$. Similarly, if $p = 0.9$ and $n = 100$, we should expect to see success 90% of the time, so $E(X) = 0.9 \cdot 100 = 90$. In general, $E(X) = np$ for binomial distributions. For a binomial random variable X with n trials, probability of success on any individual trial p, and probability of failure on any individual trial q, we can compute the expected value and variance using the following formulas.

$$\mu = \mathrm{E}(X) = np$$
$$\sigma^2 = \mathrm{Var}(X) = npq$$

? Text Exercise 4.3.4

Using the above formulas, compute the expected value and variance for the random variables: X being the number of ones rolled in 3 tosses of a fair die and Y being the number of heads in 4 flips of a fair coin. Verify that the values match what was computed in the previous text exercise.

Answer

When considering the random variable X, we have that $n = 3$, $p = \frac{1}{6}$, and $q = \frac{5}{6}$. We thus compute $\mu = \mathrm{E}(X) = 3 \cdot \frac{1}{6} = \frac{1}{2}$ and $\sigma^2 = \mathrm{Var}(X) = 3 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{12}$. These values match what was computed in the previous exercise.

When considering the random variable Y, we have that $n = 4$, $p = \frac{1}{2}$, and $q = \frac{1}{2}$. We thus compute $\mu = \mathrm{E}(Y) = 4 \cdot \frac{1}{2} = 2$ and $\sigma^2 = \mathrm{Var}(X) = 4 \cdot \frac{1}{2} \cdot \frac{1}{2} = 1$. These values again match what was computed in the previous exercise.

Necessity of Independent Trials

Binomial distributions are related to important distributions in inferential statistics, such as computing the probability of obtaining a sample with a particular proportion. Recall our discussion regarding obtaining a random sample from a large population and having 80% of them be women. The probability of this happening was significantly less with a sample size of 20 as opposed to 10 (0.4621% vs 4.3945%). These probabilities were computed using the binomial distribution. Here, we treated our random experiment as selecting an individual from a large enough population composed of equal numbers of men and women. We considered selecting a woman a success and treated $p = q = \frac{1}{2}$. In the case of a sample of size of 10, we noted that 80% of 10 is 8. And in the case of a sample size of 20, we need 16 women to get 80%. However, this fails to satisfy our definition of a binomial random variable because we do not have

the same probabilities of success and failure for each trial. When one person is chosen, that person is no longer eligible to be chosen for subsequent trials. We have fewer people to choose from and no longer equal numbers of men and women. Our trials are not independent.

We have run into this issue previously in a text exercise. When populations are huge (when the difference between an event's probability and a conditional probability related to that event is relatively small) treating the events as if they were independent will result in a value which is approximately, not exactly, correct. Since it is often much easier to compute assuming independence, this is common practice when the error would be negligible. It is difficult to define exactly how large a population must be, in general, for the assumption of independence to be reasonable. For example, if there are $1,000,000$people, exactly half of which are women, and we randomly select 2 individuals from this group, the probability that they are both women would be $\frac{500,000}{1,000,000} \cdot \frac{499,999}{999,999} \approx 0.24999975$. If we had assumed independence, that is, that each time we selected a person, there was a 50% chance it was a woman, we would have obtained $\frac{1}{2} \cdot \frac{1}{2} = 0.25$. Notice the error we get from assuming independence is quite small. On the other hand, if the population size were 6 and 3 of them were women, the assumption of independence is much less reasonable. If we randomly select 2 people from this group of 6, the probability that they are both women is $\frac{3}{6} \cdot \frac{2}{5} = \frac{1}{5} = 0.2$. Simply saying there's a 50% chance each time obtains an estimate of 0.25. Notice the error is much larger than before. If we take our population size to be even smaller, the error gets larger. In summary, if the sample we are selecting is a tiny proportion of the population, then assuming independence introduces little error; however, if we assume independence when the sample is a significant proportion of the population, then we will have large errors in our estimates. The following exercise illustrates in more detail how much error there is in different population sizes.

? Text Exercise 4.3.5

1. Consider sampling 10 people from a population composed of an equal number of men and women. We denote the outcome of such a sampling as a sequence of W and M. Determine $P(\text{WWWWWWWWMM})$ for each of population size.
   a. $N = 50$
   b. $N = 100$
   c. $N = 200$
   d. $N = 1000$

Answer

a. $P(\text{WWWWWWWWMM}) = \frac{25}{50} \cdot \frac{24}{49} \cdot \frac{23}{48} \cdot \frac{22}{47} \cdot \frac{21}{46} \cdot \frac{20}{45} \cdot \frac{19}{44} \cdot \frac{18}{43} \cdot \frac{25}{42} \cdot \frac{24}{41} \approx 0.0702\%$
b. $P(\text{WWWWWWWWMM}) = \frac{50}{100} \cdot \frac{49}{99} \cdot \frac{48}{98} \cdot \frac{47}{97} \cdot \frac{46}{96} \cdot \frac{45}{95} \cdot \frac{44}{94} \cdot \frac{43}{93} \cdot \frac{25}{92} \cdot \frac{24}{91} \approx 0.08443\%$
c. $P(\text{WWWWWWWWMM}) = \frac{100}{200} \cdot \frac{99}{199} \cdot \frac{98}{198} \cdot \frac{97}{197} \cdot \frac{96}{196} \cdot \frac{95}{195} \cdot \frac{94}{194} \cdot \frac{93}{193} \cdot \frac{100}{192} \cdot \frac{99}{191} \approx 0.09117\%$
d. $P(\text{WWWWWWWWMM}) = \frac{500}{1000} \cdot \frac{499}{999} \cdot \frac{498}{998} \cdot \frac{497}{997} \cdot \frac{496}{996} \cdot \frac{495}{995} \cdot \frac{494}{994} \cdot \frac{493}{993} \cdot \frac{500}{992} \cdot \frac{499}{991} \approx 0.09638\%$

2. Determine the $P(\text{WWWWWWWWMM})$ as if each selection were independent with $P(\text{W}) = \frac{1}{2}$ and $P(\text{M}) = \frac{1}{2}$.

Answer

$P(\text{WWWWWWWWMM}) = \left(\frac{1}{2}\right)^{10} \approx 0.09766\%$

3. Compare the value computed in each part of part 1 with the value computed in part 2 of this text exercise.

Answer

a. The difference is 0.02746%.
b. The difference is 0.01323%.
c. The difference is 0.0065%.
d. The difference is 0.0013%.

The difference in computations of these values is in the hundredths and thousandths of a percent and decreases as the population increases. We only dealt with population sizes up to 1000. In general, our populations of interest will be much larger than that so that we would expect even smaller differences. The comparison between sample size and population size is really at play down deep. Without going into the details, we share a fairly common recommendation. If the sample size is more than 5% of the population, we do not assume independence.

4.3.1: Multinomial Distributions - Optional Material

Learning Objectives

- Define multinomial random variable
- Develop the multinomial distribution

Counting Outcomes of Random Experiments

Consider the game of chess; a game can end in one of three ways: win, lose, or draw. For a pair of grandmasters, we may have an empirical estimation for the probability of each outcome based on the outcomes of previous games. If we knew they were going to be playing a set number of games n soon, we might be interested in the probability that the one player wins n_1 times, the second player wins n_2 times, and they draw n_3 times. Can we develop a random variable to handle such a task? The answer is yes; the **multinomial random variable** is a generalization of the binomial random variable. In binomial random variables, we counted the number of successful trials, which, given that we had a fixed number of trials, also set the number of failures. With three options, we must maintain counts for two of the outcomes. So, our random variable returns a coordinate pair of values.

Suppose that Magnus Carlsen and Fabiano Caruana (the two top grandmasters in June 2024) are set to play 12 games against each other in a friendly tournament. For each game, we estimate that Magnus has a 30% chance to win while Fabiano has a 25% chance to win. This leaves a 45% chance of a draw. What is the probability that of the 12 games, Magnus wins 5 games, Fabiano wins 3 games, and they draw on 4 games? Given the grandmaster status of these players, we assume that the results of previous games do not affect performances in current and future games.

We set some notation for the problem. n = 12 because 12 games are to be played, n_1 = 5 (number to be won by Magnus), n_2 = 3 (number to be won by Fabiano), n_3 = 4 (number of draws), p_1 = 0.30 (probability that Magnus wins a game), p_2 = 0.25 (probability that Fabiano wins a game), p_3 = 0.45 (probability of a draw). As mentioned above, the multinomial variable X that counts the number of wins of each player in 12 games takes on coordinate pairs of values, \($(n_1, n_2,$\) and we are interested in the probability that $n_1 = 5$ and $n_2 = 3$, $P(X = (5, 3))$.

With 12 games and 3 possible outcomes for each game, considering every possible sequence of 12 outcomes is out of the question. We would have $3^{12} = 531,441$ sequences to consider. Hopefully, we can build on our understanding of the binomial random variable. Recall that the probability of a particular sequence of outcomes of all the trials depended on the total number of successes and failures. The order in which they occurred did not matter. This probability was $p^{x_j} q^{n-x_j}$. We then counted the number of ways that a number of successes and failures could happen, ${}_nC_{x_j}$, which led to our probability computation of ${}_{n}C_{x_j}\, p^{x_j} q^{n-x_j}$.

A similar line of reasoning will help us develop a probability distribution function for multinomial variables. Just as with binomial random variables, the probability of a particular sequence of outcomes depends on the values of n_1, n_2, and $n_3 = n - n_1 - n_2$. We arrive at the probability computation $p_1^{n_1} p_2^{n_2} p_3^{n_3}$. The only issue remains to count the number of such sequences that have given n_1 and n_2 values. Here, we refer to the optional material in chapter 3 : distinguishable permutations. We have three outcomes that we are assigning to particular trials, and the order in which a trial is assigned to one of these outcomes does not matter. We can, therefore, count the number of sequences that have given n_1 and n_2 values with this computation: $\frac{n!}{n_1! n_2! n_3!}$. We conclude that the probability distribution function for a multinomial random variable X with 3 outcomes and n trials.

$$\begin{aligned} P(X = (n_1, n_2)) &= \frac{n!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3} \\ &= \frac{n!}{n_1! n_2! (n - n_1 - n_2)!} p_1^{n_1} p_2^{n_2} p_3^{n - n_1 - n_2} \end{aligned}$$

We can answer our original question in the context of chess: $P(X = (5, 3))$.

$$\begin{aligned} P(X = (5, 3)) &= \frac{12!}{5!3!4!} (0.30)^5 (0.25)^3 (0.45)^4 \\ &\approx 27,720 \cdot 0.00243 \cdot 0.01563 \cdot 0.04101 \\ &\approx 4.3159\% \end{aligned}$$

Text Exercise 4.3.1.1

Suppose that Magnus and Fabiano decide that 12 games are too many and reduce it to just 4 games. Produce the probability distribution for the multinomial random variable X that counts each of their wins.

Answer

Since there are three outcomes that we are interested in rather than just two with binomial random variables, we have many more options to consider, 15 options in fact.

Table 4.3.1.1 Probability distribution for the random variable X

$X=(n_1,n_2)$	$P(X=(n_1,n_2))$	$X=(n_1,n_2)$	$P(X=(n_1,n_2))$
$(0,0)$	$\frac{4!}{0!0!4!}(0.30)^0(0.25)^0(0.45)^4\approx 4.1\%$	$(1,0)$	$\frac{4!}{1!3!0!}(0.30)^1(0.25)^3(0.45)^0\approx 1.8$
$(0,1)$	$\frac{4!}{0!1!3!}(0.30)^0(0.25)^1(0.45)^3\approx 9.2\%$	$(1,1)$	$\frac{4!}{2!0!2!}(0.30)^2(0.25)^0(0.45)^2\approx 10.$
$(0,2)$	$\frac{4!}{0!2!2!}(0.30)^0(0.25)^2(0.45)^2\approx 7.3\%$	$(1,2)$	$\frac{4!}{2!1!1!}(0.30)^2(0.25)^1(0.45)^1\approx 12.$
$(0,3)$	$\frac{4!}{0!3!1!}(0.30)^0(0.25)^3(0.45)^1\approx 2.8\%$	$(2,0)$	$\frac{4!}{2!2!0!}(0.30)^2(0.25)^2(0.45)^0\approx 3.3$
$(0,4)$	$\frac{4!}{0!4!0!}(0.30)^0(0.25)^4(0.45)^0\approx 0.39\%$	$(3,0)$	$\frac{4!}{3!0!1!}(0.30)^3(0.25)^0(0.45)^1\approx 4.8$
$(1,0)$	$\frac{4!}{1!0!3!}(0.30)^1(0.25)^0(0.45)^3\approx 18.94\%$	$(3,1)$	$\frac{4!}{3!1!0!}(0.30)^3(0.25)^1(0.45)^0\approx 2.7$
$(1,1)$	$\frac{4!}{1!1!2!}(0.30)^1(0.25)^1(0.45)^2\approx 18.23\%$	$(4,0)$	$\frac{4!}{4!0!0!}(0.30)^4(0.25)^0(0.45)^0\approx 0.8$
$(1,2)$	$\frac{4!}{1!2!1!}(0.30)^1(0.25)^2(0.45)^1\approx 10.13\%$		

Multinomial random variables can extend to counting many more outcomes. We conclude this section by generalizing the multinomial random variable where we count k outcomes. The probability distribution function for a multinomial random variable X with k outcomes and n trials is given below.

$$P(X=(n_1,n_2,\ldots,n_{k-1}))=\frac{n!}{n_1!n_2!\ldots n_k!}p_1^{n_1}p_2^{n_2}\ldots p_k^{n_k}$$

Note that the binomial distribution is a special case of the multinomial distribution when $k=2$.

4.4: Continuous Probability Distributions

Learning Objectives

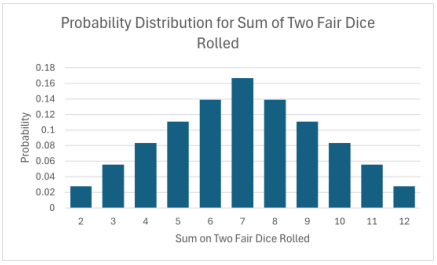
- Define the meaning of a continuous random variable probability distribution and its associated probability density function
- Use graphs to represent continuous random variables' probability distribution
- Connect area under the probability density function to probability measures for a continuous random variable
- Find area/probability measures for distributions with basic shapes

Review and Preview

We have introduced the concept of probability distributions for random variables: a distribution that represents all possible outcomes of a random variable and the associated probabilities for each. For example, we examined the discrete random variable of the sum of two rolled dice. The outcomes were sums of value 2 through 12, and the probability of each is given in the table below. A table is one way to represent the probability distribution; another is to produce a bar graph to have a pictorial representation of the distribution. We noted that the sum of the probabilities must total $1 = 100\%$ to have a complete probability distribution.

Table 4.4.1: Probability distribution of the sum of two fair dice in graphical and tabular formats

X: Sum on Two Dice Rolled	**Probability $P(X = x_j)$**	**Graphic Representation**
2	$\frac{1}{36}$	
3	$\frac{1}{18}$	
4	$\frac{1}{12}$	
5	$\frac{1}{9}$	
6	$\frac{5}{36}$	
7	$\frac{1}{6}$	
8	$\frac{5}{36}$	
9	$\frac{4}{36}$	
10	$\frac{3}{36}$	
11	$\frac{2}{36}$	
12	$\frac{1}{36}$	
Total:	$\sum P(x_j) = \frac{36}{36} = 1.0000 = 100\%$	



Another critical concept in the above example was that the random variable was discrete. Each outcome could be listed, and the probability of each outcome was determined. Other examples include the random variable "number of days adults exercise per week" or the random variable "amount of change in teenagers' pockets."

Next, we discussed finding the mean ("expected value"), variance, and standard deviation measures from our discrete probability distribution tables. We saw how the computation concepts of grouped data (Sections 2.8 and 2.9) are used to find these measures in our probability distributions.

We also discussed in Section 4.3 a unique collection of discrete probability distributions called Binomial Distributions, distributions whose random variable is the number of successes in a given situation. If we have a well-defined success and failure in the situation, a fixed number of independent trials, and a fixed probability of success in the trials, then the probability distribution for the binomial situation is reasonably easy to construct.

Once we have the probability distribution table for a discrete random variable, we can use that information, along with our probability rules, to determine probability measures in relation to any outcomes of interest.

Now, we turn our focus to probability distributions of continuous random variables. Recall the example from Section 4.1 about the random variable of "the time (in seconds) it takes both dice in a two-dice roll to come to a complete stop after one die leaves our hand." We can no longer get accurate probability measures from a table listing outcomes and associated probabilities as in the discrete cases above. For example, there are always possible outcomes on a continuous variable between other values. Although we might build an estimated probability distribution table using intervals on the continuous random variable, doing so causes us to lose information about the distribution of the variable. We must use a different approach to maintain reasonable accuracy in dealing with continuous random variables.

Continuous Probability Distributions

Recall that a continuous variable can take on any numerical value in an interval of real numbers; in particular, another value exists between any two possible values. Examples of such variables included height, weight, ounces of water consumed, time elapsed, age, amount of electricity consumed, and many more. We must be aware that even though another height measure exists between any two heights, we measure using some chosen discrete scale, such as to the nearest inch. This rounded height measure does not make the variable discrete, the variable is still continuous. We use this information in our following theory on probability distributions on continuous variables.

A **continuous random variable probability distribution** assigns probability to an interval of values of the continuous random variable. For example, the probability distribution on the continuous variable height should give us the probability of randomly selecting a person whose height is between 5 feet and 6 feet; it should also assign a probability to any other interval of choice. This is where we move away from histograms and relative frequency tables that have specifically chosen intervals for the classes.

In Section 2.7, we demonstrated the use of a continuous mathematical function that matches the shape of a histogram graphic. Our example from that section is given below.

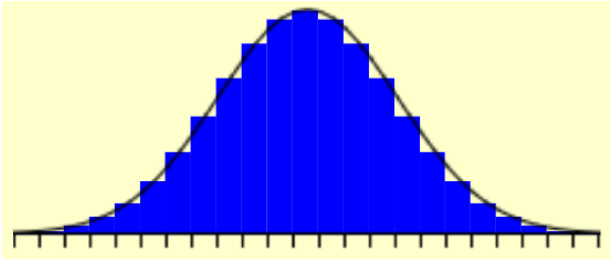


Figure 4.4.1: Histogram with fitted curve

In modeling a continuous variable's distribution, we produce a curve that matches the behavior of the various classes in the histogram. If we move to more and more narrow class intervals, the variable will follow a function's curve. Another example is given below, in which we demonstrate the curve matching to a distribution that is positively skewed.

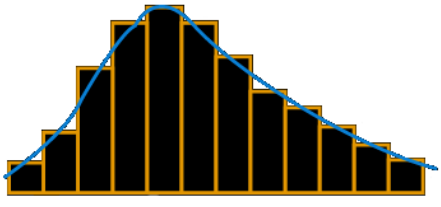


Figure 4.4.2: A second histogram with fitted curve

To have a probability distribution, our variable's distribution and the curve fitting pf that distribution must be tied to probability (which is closely related to relative frequency). Relative frequency histograms tend to lose meaning as the width of their classes decreases, as pictured below on the left column in Figure $4.4.3$. For example, recalling that this data set represents the heights of people, notice that a little over 20% of people are 68 inches tall when their height is rounded to the nearest inch. If we instead measured height to the nearest tenth of an inch, we see approximately 2% are 68.1 inches tall, 2% are 68.2 inches tall, and so on. As we get more precise with our measurements, the proportion of people in any particular class gets smaller; hence, the relative frequencies go to 0. If the heights of the bars are the relative frequencies, then the picture degenerates. A way to overcome this issue is to represent the relative frequencies as areas instead of as heights. This is shown again below.

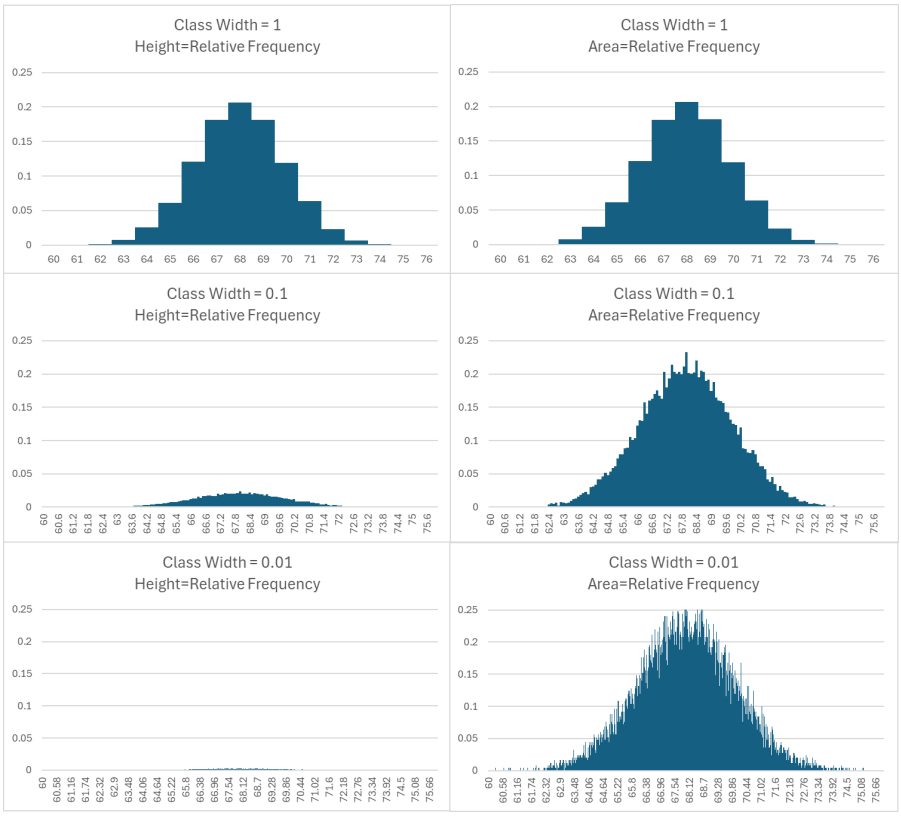


Figure $4.4.3$: Probabilities of classes as class width decreases (height of bar on left and area of bar on the right)

The curves that fit the area graphics are called **probability density functions** (PDFs for short.) The function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ from the symmetric bell-shaped curve (commonly called the normal distribution) is the probability density function for the normal curve with a population mean of μ and standard deviation of σ. The curves are called density functions because the curve values are not directly probability measures but are measures of the denseness of probability. To find probability values, we measure the area under the density function values over an interval of values. We build regions under the probability density curve whose area measures equate to probability measures. This connection and its use will become more evident in the following sections.

All probability density functions for continuous random variables will always have three key features.

1. The domain of the curve (even if the continuous random variable has a smaller domain) can be all real numbers (in interval notation: $(-\infty, \infty)$).
2. The function values $f(x)$ for the density function will always be non-negative values; that is $f(x) \geq 0$ for all values of the continuous random variable x.
3. The total area under the curve is equal to $1 = 100\%$, and the area under the curve over an interval $a \leq x \leq b$ of the continuous variable will produce the probability measure $P(a \leq x \leq b)$.

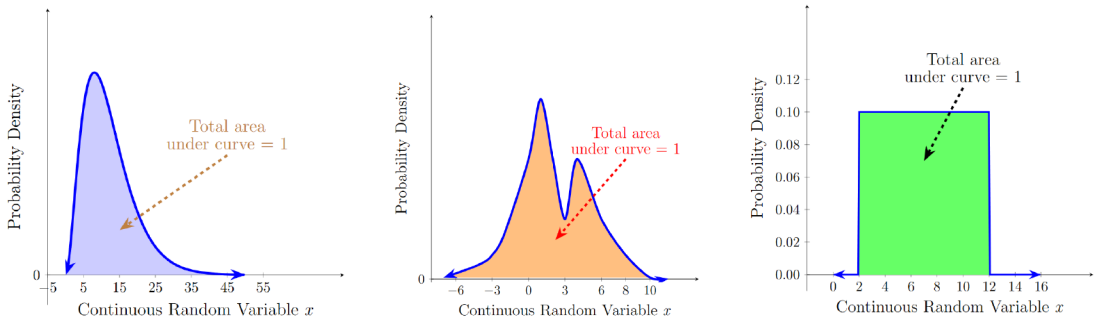


Figure 4.4.4: Examples of probability density functions

The horizontal scale will depend on the random variable being investigated. For example, the random variable represented by the left density function in Figure 4.4.4 has its most commonly occurring outcomes between 0 and 50, the random variable of the center density function has most outcomes between -7 and 10, the random variable of the right density function has outcomes between 2 and 12. These horizontal scales are very important to the meaning of each random variable and that variable's distribution and should be included. We also note that, at times, vertical axis scaling will not be explicitly given when working with PDF graphs (compare the left two above with the right one); in general, this should not cause us concern provided we know the curve is a PDF satisfying our three requirements.

We also briefly note that these probability density functions approximate probability measures for discrete cases due to the many mathematical benefits of such curves. For example, if dealing with a binomial distribution situation in which the number of trials is large, say 500 trials, instead of building a binomial distribution table of variable values from 0 to 500--a huge table to work with--we can approximate that distribution with a single appropriate density function. This allows us to use functions instead of building a large table to examine the distribution.

Let's examine this connection between area and probability with continuous variable probability distributions.

Probability Measures from Continuous Probability Distributions

We first examine graphs of probability distributions and answer some questions concerning those distributions. For example, we might be given the graph below as a proposed probability distribution of a continuous random variable x.

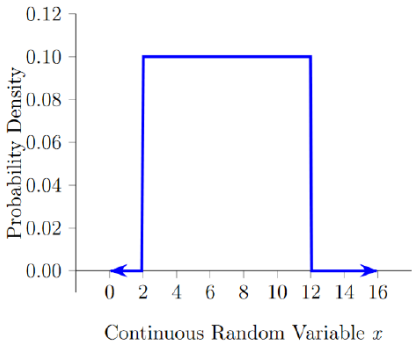


Figure 4.4.5: Example Probability Density Function

Notice for all values $2 \leq x \leq 12$; the graph shows a density function value of $PDF(x) = 0.10$, and for all other real number values of x, we have $PDF(x) = 0$. This graph implies that the continuous variable x only has possible outcomes between 2 and 12; all other real values are "impossible" outcomes since their probability density is 0. In such graphs, we focus on those intervals of variable values where $PDF(x) \neq 0$.

We should check that the three requirements of a $PDF(x)$ are met in this graphic. First, the curve's domain is all real numbers, as implied by the arrows at the end of the blue curve. Next, for all x, we see that $f(x) \geq 0$. Finally, we notice the rectangular region between the curve and the x-axis over the interval $2 \leq x \leq 12$. The width of this rectangle is $12 - 2 = 10$ units, and the height of this rectangle is a probability density measure of 0.10 units. The area calculation finds the enclosed area between the curve and x-

axis on the rectangle: $\text{Area} = \text{base} \cdot \text{height} = 10 \cdot \frac{1}{10} = 1 = 100\%$. We have a total probability measure of $1.00 = 100\%$ in this curve's area measure.

Even if we don't know the specific real-life context, this curve mathematically represents the probability distribution of some continuous random variable x. This graphic will allow us to find probability measures for different interval values; again, we focus only on intervals in which the PDF is non-zero to eliminate unnecessary work involving impossible outcomes for the variable.

For this variable x, with the given probability distribution shown above, we may wonder what the probability of randomly selecting outcomes over the interval $7 < x < 10$ would be; that is, we wish to determine $P(7 < x < 10)$. To illustrate, we can color the area within this distribution that coincides with the x values of the interval.

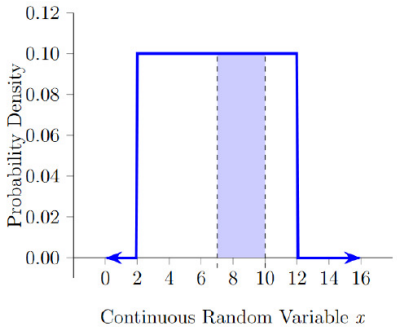


Figure 4.4.6: Finding $P(7 < x < 10)$

We notice that our shaded region is rectangular. The area of this shaded rectangle is the measure of the probability. The width of this rectangle is $10 - 7 = 3$ units in the continuous variable, and the height of this rectangle is a probability density measure of 0.10 units. The shaded area is again found by calculating the area of the rectangle:

$$\begin{aligned} \text{Area} &= \text{base} \cdot \text{height} \\ &= 3 \cdot \frac{1}{10} \\ &= 0.30 = 30\%. \end{aligned}$$

In this distribution, $P(7 < x < 10) = 0.30 = 30\%$. If we randomly select an outcome in this situation, then 30% of the time, we would expect to see an outcome between 7 and 10. Stated equivalently, 30% of outcomes in this x–variable's distribution are between 7 and 10.

? Text Exercise 4.4.1

Using our distribution of Figure 4.4.5, find the following probability measures.

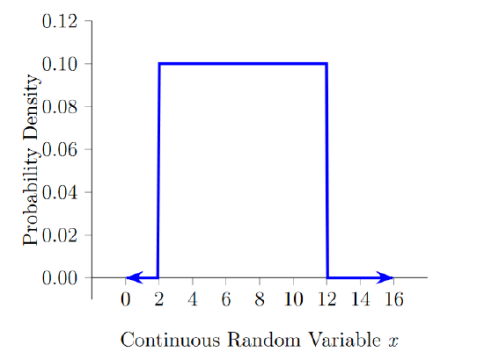


Figure 4.4.5: Replication of previous probability distribution

1. Determine $P(x \leq 8.5)$.

Answer

We shade the region under the density curve over the variable's interval $x < 8.5$.

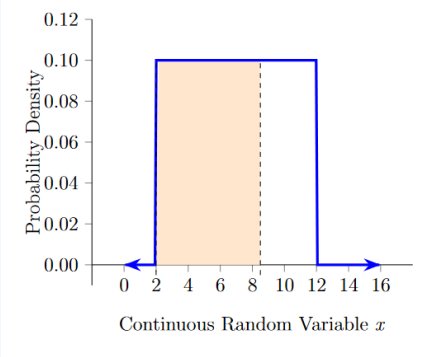


Figure 4.4.7 Finding $P(x < 8.5)$

Our shaded region is a rectangle with width $8.5 - 2 = 6.5$ and height of 0.10. So,

$$\begin{aligned} P(x < 8.5) &= \text{area of the region} \\ &= 6.5 \cdot 0.10 \\ &= 0.65 = 65\%. \end{aligned}$$

About 65% of this continuous variable's outcomes are less than 8.5 units.

2. Determine $P(x > 8.5)$.

Answer

We will take two approaches to make a critical point. Using the same approach, we shade under the density curve over the variable's interval $x > 8.5$.

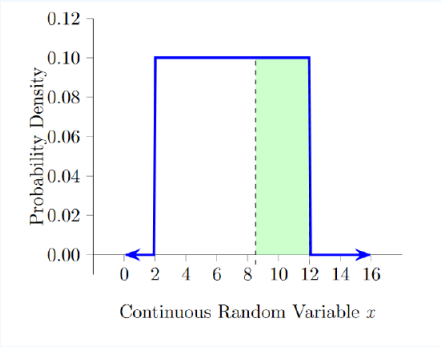


Figure 4.4.8 Finding $P(x > 8.5)$

Our region is a rectangle with width $12 - 8.5 = 3.5$ and height of 0.10. So,

$$\begin{aligned} P(x \geq 8.5) &= \text{area of the shaded region} \\ &= 3.5 \cdot 0.1 \\ &= 0.35 = 35\% \end{aligned}$$

We might show or not show solid or dashed vertical boundary lines on our regions; inclusion or exclusion will not make a measurement difference in the area.

3. Determine $P(2.75 < x < 5.5)$.

Answer

We shade under the density curve over the variable's interval $2.75 < x < 5.5$.

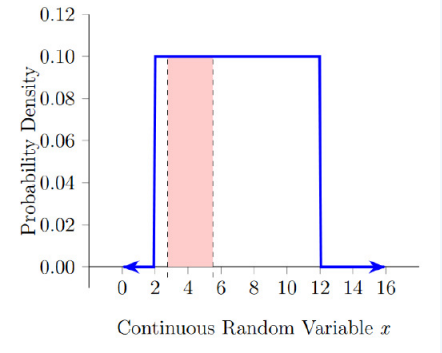


Figure 4.4.9 Finding $P(2.75 < x < 5.5)$

Our shaded region is a rectangle with width $5.5 - 2.75 = 2.75$ and height of 0.10

$$\begin{aligned} P(2.75 < x \leq 5.5) &= \text{area of the region} \\ &= 2.75 \cdot 0.10 \\ &= 0.275 = 27.5\% \end{aligned}$$

About 27.5% of this continuous variable's outcomes are between 2.75 and 5.5 units.

The above examples and exercises were relatively straightforward since the regions of interest were always rectangles. Naturally, not all continuous random variables will have this same distribution shape.

Note: Strict and Non-Strict Inequalities

Here, we emphasize a crucial point. In our work, we make no distinction in area measures from regions formed on strict inequalities, such as $<$ or $>$, on a continuous random variable and other inequalities, such as $\leq$ or $\geq$. With continuous distributions, there is 0 area under the curve over a single value, that is, technically $P(x = a) = 0$ for any single outcome a. Therefore, the area measure of regions such as $P(x < a)$ is the same as for $P(x \leq a)$.

Due to this, when dealing with regions under continuous probability distribution functions, strict inequalities can be used interchangeably with non-strict inequalities. In our graphics of regions, we may or may not show dark or dashed vertical boundary lines on our regions; inclusion or exclusion will not make a measurement difference in the area.

We also remind ourselves that there is a difference, in general, between the use of strict and non-strict inequalities in discrete distribution probabilities discussed in earlier sections of this chapter. This demonstrates another reason why it is important to know if the random variable being analyzed is continuous or discrete.

Now, let us examine a different continuous probability distribution.

Note: Pertinent Common Area Formulas

Rectangle: $\text{base} \cdot \text{height}$

Triangle: $\frac{1}{2}\text{base} \cdot \text{height}$

Trapezoid: $\frac{\text{base}_1 + \text{base}_2}{2}\text{height}$

Text Exercise 4.4.2

Suppose the following continuous variable distribution is given. Answer the following questions concerning this distribution.

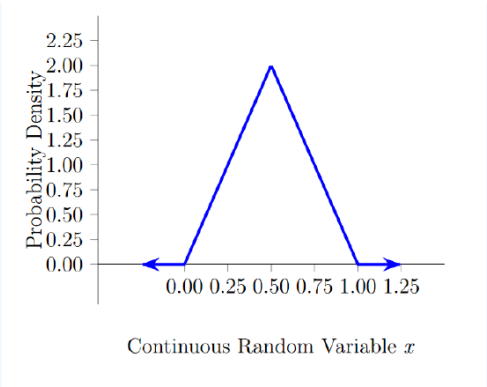


Figure 4.4.10: Another continuous variable distribution

1. Determine if this graph can be a probability density function for a continuous variable.

Answer

We notice for all values $0 \leq x \leq 1$, the graph shows a changing density value, increasing from 0 to 2 and then decreasing back to 0. For all other real number values of x, we have $PDF(x) = 0$. This graph implies that the continuous variable x only has possible outcomes between 0 and 1; all other real values are "impossible" outcomes since their probability density is 0.

We also check that the three requirements of a $PDF(x)$ are truly met in this graphic. Notice that the domain of the curve is all real numbers, as implied by the arrows at the end of the blue curve. Next, for all x, we see that $f(x) \geq 0$. Finally, we notice a triangular region between the curve and the x-axis over the interval $0 \leq x \leq 1$. The base of this triangle is $1 - 0 = 1$ unit in the continuous variable, and the height of this rectangle is a probability density measure of 2.00units. So the enclosed area between the curve and x-axis is found by the area calculation on triangles:

$$\begin{aligned} \text{Area} &= \frac{\text{base} \cdot \text{height}}{2} \\ &= \frac{1 \cdot 2}{2} \\ &= 1 = 100\% \end{aligned}$$

Thus we have total probability measure of $1.00 = 100\%$in this curve's area. This analysis establishes that this curve does represent a probability density function for some continuous variable.

2. Determine $P(x \leq 0.50)$.

Answer

To find $P(x \leq 0.50)$, we shade under the density curve over the variable's interval $x \leq 0.50$.

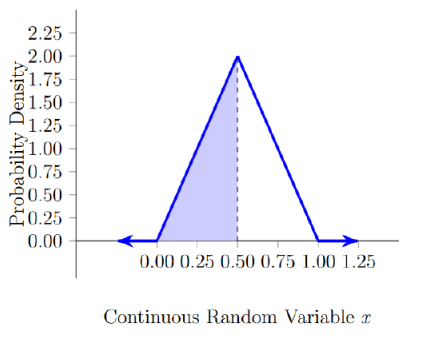


Figure 4.4.11: Finding $P(x \leq 0.50)$

We might easily notice that our shaded region is half of the total region. This implies that $P(x \leq 0.50) = 0.50 = 50\%$.

To check this, we notice that our shaded region is a triangle with base $0.50 - 0.00 = 0.50$and height of 2.00.So,

$$\begin{aligned} P(x \leq 0.50) &= \text{area of the shaded region} \\ &= \frac{0.50 \cdot 2.00}{2} \\ &= 0.50 = 50\%. \end{aligned}$$

About 50% of this continuous variable's values are at most 0.50 units on the continuous scale. We note that many similar numbers are involved in this problem; often, we must focus more on the meaning of the values we use as we compute instead of the actual values themselves.

3. Determine $P(x \geq 0.75)$.

Answer

To find $P(x \geq 0.75)$, we again shade the related region under the density curve over the variable's interval $x \geq 0.75$.

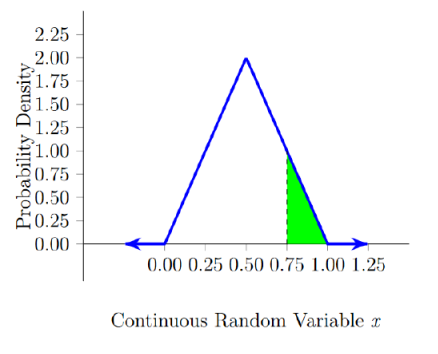


Figure 4.4.12 Finding $P(x \geq 0.75)$

4. Determine $P(0.25 \leq x \leq 0.50)$.

Answer

To find $P(0.25 \leq x \leq 0.50)$, we shade under the density curve over the variable's interval $0.25 \leq x \leq 0.50$.

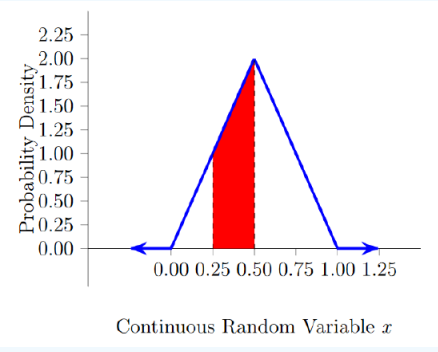


Figure 4.4.13 Finding $P(0.25 \leq x \leq 0.50)$

We will find the area of this shaded region using two approaches. It is not about one way of thinking but finding a reasonable way to determine the shaded region's area.

For our first approach, we recall the total area under all the density functions is $1.00 = 100\%$. We also notice that the shaded region is surrounded by two triangles (white regions in the graphic). If we subtract the area of the two white triangles from the total area of 1.00,we will be left with the area of the red region. That is,

$$\begin{aligned} P(0.25 \leq x \leq 0.50) &= 1.00 - (\text{area of left white triangle} + \text{area of right white triangle}) \\ &= 1.00 - \left(\frac{0.25 \cdot 1.00}{2} + \frac{0.50 \cdot 2.00}{2} \right) \\ &= 1.00 - (0.125 + 0.50) \\ &= 1.00 - 0.625 = 0.375 = 37.5\%. \end{aligned}$$

So 37.5% of this continuous variable's values are between 0.25 and 0.50 units.

As a different approach, we might notice that the red-shaded region is a trapezoid, and the area of a trapezoid is found by the average of the parallel sides (commonly called the trapezoid bases) multiplied by the distance between the parallel sides (commonly called the height of the trapezoid). Using knowledge of trapezoids,

$$\begin{aligned} P(0.25 \leq x \leq 0.50) &= \text{area of red trapezoid} \\ &= \frac{1.00 + 2.00}{2} \cdot (0.50 - 0.25) \\ &= \frac{3}{2} \cdot 0.25 \\ &= 0.375 = 37.5\%. \end{aligned}$$

Although we found the area using a different approach, we see that 37.5% of this continuous variable's values are between 0.25 and 0.50 units.

As long as we have a probability distribution on a continuous variable with an appropriate probability density function, we can answer any probability question for that variable by finding the area of the related regions. Since we are naturally curious, our minds wonder what happens if our regions of interest are not always simple geometric figures. We examine such distributions at times in the following two text sections.

Summary

This section has connected probability distribution graphs on continuous random variables, probability density functions, and areas under probability density functions. Specifically, to find the probability of an interval of values for a continuous random variable, we must find the area under the related probability density function over the interval of interest. The following section will examine some of statistical analysis' most common continuous probability distributions.

4.5: Common Continuous Probability Distributions

Learning Objectives

- Sketch graphs of continuous random variable distributions based on a given description geometric, $(\text{normal}, t, \text{and } \chi^2)$
- Use basic geometry to determine probability measures in certain continuous random variable distributions
- Know key properties of other common continuous random variable distributions $(\text{normal}, t, \text{and } \chi^2)$
- Relate various regions of a continuous random variable's distribution with each other, specifically, relate any region to left-tail region(s)

Review and Preview

In Section 4.4, we established the connection between area measures of regions under the probability density function of a continuous random variable and the probability of certain outcome intervals for that variable; namely, we showed that the area of the region over an interval is the probability value. We examined a few examples of probability distributions. In this section, we name and explore key properties of some of the most commonly used probability density functions in statistical work. We will end by examining how certain regions within our distributions can directly relate to other regions within our distributions using some basic geometric reasoning.

Distributions: Shapes from Basic Geometry

Our last section examined exercises involving two random variables with different distributions. As shown below in Figure 4.5.1, one is rectangular, and the other is triangular.

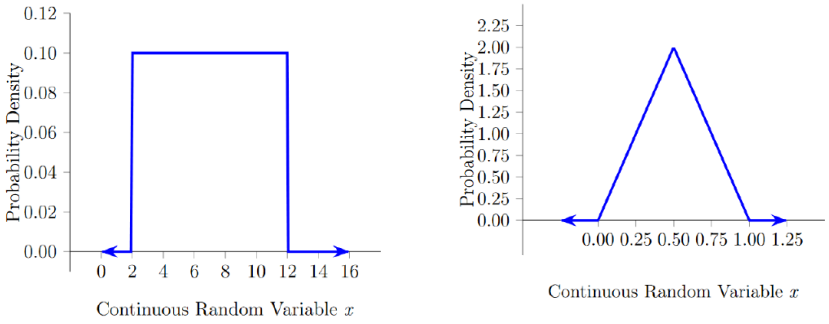


Figure 4.5.1: Two distributions with shapes from basic geometry

These two distribution shapes can be found in professions such as finance, ecology, business, education, and others. The rectangular-shaped distribution is a **uniform probability distribution** with a similar meaning to the uniform distribution on discrete random variables (such as a fair dice roll). Rather than each possible outcome having the same probability, every possible outcome has the same probability density. As a result, every interval of possible outcomes of the same width has an equal probability of occurring. For example, in the uniform distribution in Figure 4.5.1, an outcome between 2 and 3 is equally as likely as an outcome between 7 and 8. In the picture above, every interval of equal length contained in $[2, 12]$ is equally likely. Naturally, there is not just one uniform distribution, but all uniform distributions on continuous variables form a family with some common properties. One such property is that all uniform distributions are symmetric. Using the idea of a "balance point for the center of mass," the mean of the uniform distribution above is at $\mu = 7$, the midpoint of the interval $[2, 12]$. The median is also at 7 since 50% of the rectangle's area is below that $x-$value and 50% is above. In general, the mean and median of any uniform distribution will always be this midpoint value.

The triangular-shaped distribution on the right is commonly called a **triangular probability distribution**. Although the shape of the given triangular distribution in Figure 4.5.1 above is symmetric, this cannot be said of all triangular distributions. Again, we have a family of triangular distributions when probability density function curves form a triangular shape. With this symmetric triangular probability distribution in Figure 4.5.1, we can see the "balance point for the center of mass" to be at $\mu = 0.50$ on the horizontal scale and the median to also be at 0.50. In some triangular distributions, determining the mean and the median is not as easy without more advanced skills in mathematics. Although we will not always determine these key statistical measures for every

distribution, it is important to realize that these, and many of our summary statistics discussed in Chapter 2, exist in probability distributions.

There are many distributions with shapes from basic geometry, such as semi-circles or trapezoids. But as discussed in the last section, the total area under the curve must always total $1 = 100\%$ and the density function heights must always be non-negative $(f(x) \geq 0)$. If we can find the area of regions under the density functions over an interval, we can interpret those areas as probability values.

Text Exercise 4.5.1

As a random variable, data on the daily growth of the height of wheat plants during a particular stage of development is believed to be uniformly distributed between $\frac{1}{2} = 0.5$ and $\frac{5}{4} = 1.25$ inches. Answer the following questions about this variable in the context of wheat growth.

1. Sketch a graph of the probability distribution on the wheat's growth height, including appropriate labeling of both axes.

Answer

Based on the information given, and choosing to work in decimal representation of values, we build the probability distribution graph by placing a horizontal line segment above our random variable's horizontal axis over the interval $[0.5, 1.25]$and horizontal line segments on the x-axis for all other real numbers of the scale. Knowing that the total area under the distribution curve must equal $1 = 100\%$, and that our non-zero probability interval has a width of $1.25 - 0.5 = 0.75$, the height of the rectangular portion of the uniform distribution must satisfy

$$\text{height} = \frac{\text{area}}{\text{base}} = \frac{1}{0.75} = \frac{4}{3} \approx 1.3333.$$

With full labeling of the axes, we produce the following graph of the probability distribution:

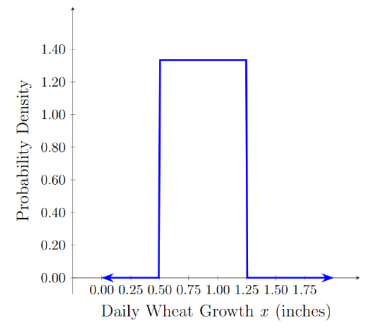


Figure 4.5.2 Probability distribution for the random variable daily wheat growth

2. What is the distribution's expected value (mean) and median?

Answer

As discussed above, the interval's midpoint $[0.5, 1.25]$ produces both the expected value and the median in uniform distributions, $\frac{0.5+1.25}{2} = 0.875$ inches. When known within a context, units should be included where appropriate to bring meaning to reported values.

3. Find the probability that a randomly selected wheat plant will grow at least 1.0 inch in a given day.

Answer

After shading above the desired interval in our graphic, to find $P(x \geq 1.0)$, we see the area of the shaded region.

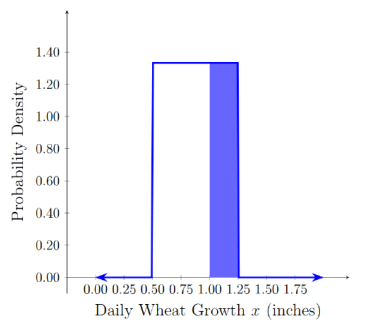


Figure 4.5.3 Finding $P(x \geq 1.0)$

This shaded region within the uniform distribution is a rectangle with a base length of 0.25 inches and height with a density value of $\frac{4}{3} \approx 1.3333$, so $P(x \geq 1.0) \approx 0.25 \cdot 1.3333 \approx 0.3333$. We have a 33.33% probability of randomly selecting a wheat plant that will grow more than 1 inch in a given day.

4. What proportion of wheat plants are expected to grow between 0.6 to 0.75 inches in a given day? In the uniform distribution, find $P(0.6 \leq x \leq 0.75)$.

Answer

After shading above the desired interval in our graphic, to find $P(0.6 \leq x \leq 0.75)$, we see the area of the shaded region.

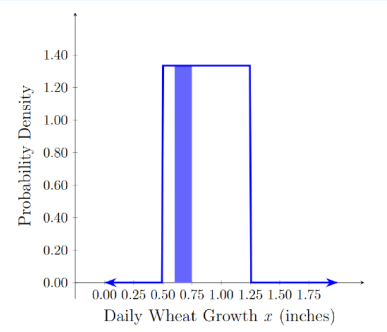


Figure 4.5.4 Finding $P(0.6 \leq x \leq 0.75)$

This shaded region is a rectangle with base length of $0.75 - 0.6 = 0.15$ inches and height again with density value of $\frac{4}{3} \approx 1.3333$, so $P(0.6 \leq x \leq 0.75) \approx 0.15 \cdot 1.3333 = 0.2$. This means 20% of wheat plants are expected to grow between 0.6 to 0.75 inches in a given day.

? Text Exercise 4.5.2

An oil company has data showing that an old oil field in central Kansas produces between 1000 and 2500 barrels of oil every day. Their data indicates the production distribution is triangular, with the most common daily production at 1500 barrels. Answer the following questions about this oil field:

1. Sketch a graph of the probability distribution for the oil field's production, including appropriate labeling of both axes.

Answer

Based on the information given and choosing to work in decimal representation of values, we build the probability distribution graph by placing a triangular shape above our random variable's horizontal axis over the interval $[1000, 2500]$ with the peak of the triangle occurring at 1500. Also, we have horizontal line segments on the x-axis for all other real numbers on the scale. We recall that triangle area is found by $\text{area} = \frac{\text{base} \cdot \text{height}}{2}$, so with basic algebraic manipulation we have $\text{height} = \frac{\text{area} \cdot 2}{\text{base}}$. Knowing that the total area under the distribution curve must equal $1 = 100\%$ and that our non-zero probability interval has a width of $2500 - 1000 = 1500$, the highest point of the triangle-shaped distribution must satisfy

$$\text{height} = \frac{1 \cdot 2}{1500} = \frac{2}{1500} = \frac{1}{750} \approx 0.001333.$$

With full labeling of the axes, we produce the following graph of the probability distribution.

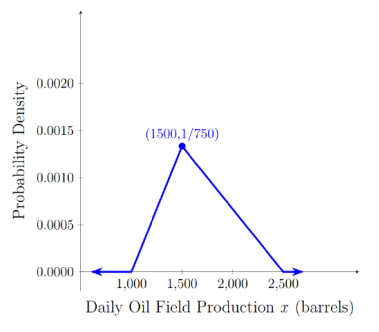


Figure 4.5.5 Probability distribution for daily oil field production

2. Find the probability that a randomly selected day will result in less than 1500 barrels of production.

Answer

After shading above the desired interval in our graphic, to find $P(x < 1500)$, we see the area of the shaded region.

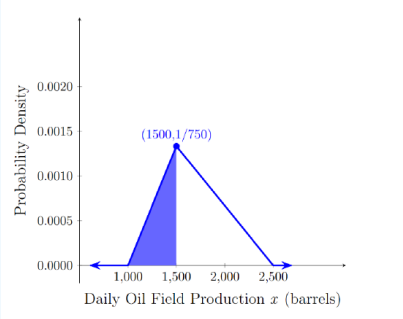


Figure 4.5.6 Finding $P(x < 1500)$

The area of the shaded region is a simple triangle. This triangle has a base length of 500 barrels and height with a density value of $\frac{1}{750} \approx 0.001333$, so $P(x < 1500) \approx \frac{500 \cdot 0.001333}{2} \approx 0.3333$. We have a 33.33% probability of randomly selecting a day in which this oil field will produce less than 1500 barrels.

3. Find the probability that a randomly selected day will result in less than 2000 barrels of production.

Answer

After shading above the desired interval in our graphic, to find $P(x < 2000)$, we see the area of the shaded region.

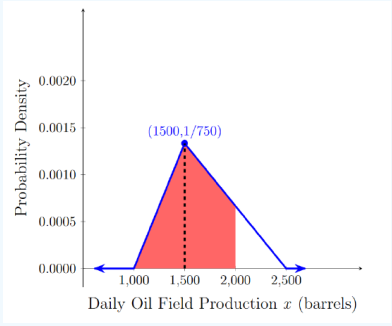


Figure 4.5.7 Finding $P(x < 2000)$

The shaded region is not a triangle, but we might notice that the region is triangular from 1000 to 1500. The rest of the region between 1500 and 2000 is a trapezoid. We can find the area of those two regions and add them together to get the total area.

We might approach this a bit easier by use of our complement rule on probabilities, that $P(x < 2000) = 1 - P(x \geq 2000)$, and noticing the white region associated with $P(x \geq 2000)$ is a simple right triangle. In that triangle, we see the base length of $2500 - 2000 = 500$ barrels. The height, however, is a bit more challenging to determine. This can be done in multiple ways (such as using the slope concept of lines with the graph scale). Let us develop the linear function for the line on the right side of the probability distribution to produce other density values if needed in future work. This also demonstrates that knowing the density function's mathematical formula can be helpful.

Using the point-slope approach, we note that the slope can be determined from the two points $(1500, \frac{1}{750})$ and $(2500, 0)$.

$$\text{slope} = \frac{\frac{1}{750} - 0}{1500 - 2500} = -\frac{1}{750,000}$$

Using our point slope-form of a line and choosing the point $(2500, 0)$ to work with, we have the linear function

$$y = -\frac{1}{750,000}(x - 2500)$$

We see the density value is given as

$$y = -\frac{1}{750,000}(2000 - 2500) = \frac{1}{1500} \approx 0.0006667.$$

Now that we know the height of our white triangular region in our graphic, we can compute that triangle's area:

$$\begin{aligned} P(x < 2000) &= 1 - P(x \geq 2000) \\ &= 1 - \frac{500 \cdot \frac{1}{1500}}{2} \\ &= 1 - \frac{1}{6} = \frac{5}{6} \approx 0.8333 \end{aligned}$$

We have an 83.33% probability of randomly selecting a day in which this oil field will produce less than 2000 barrels.

4. What proportion of days will the oil field be expected to produce between 1750 and 2000 barrels? That is, find $P(1750 \leq x \leq 2000)$ in the triangular distribution.

Answer

To find $P(1750 \leq x \leq 2000)$, we quickly shade our triangular distribution appropriately and notice the region is a trapezoid.

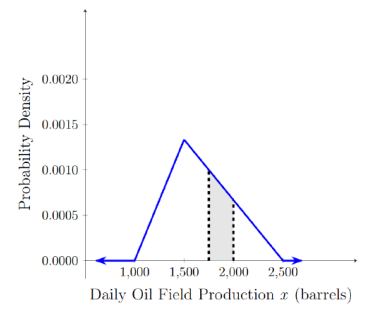


Figure 4.5.8 Finding $P(1750 \leq x \leq 2000)$

To find the area of the trapezoid, we need both density values associated with productions of 1750 barrels and of 2000 barrels. Using our work from part 3 above, we know the density associated with 2000 barrels is $\frac{1}{1500} \approx 0.0006667$ and, using our developed linear function, the density associated with 1750 barrels is $-\frac{1}{750,000}(1750 - 2500) = -\frac{1}{750,000} \cdot -750$

$= \frac{1}{1000} = 0.001$. These density values are the lengths of our parallel sides of the trapezoid and the width of the trapezoid is the interval width of $2000 - 1750 = 250$. Our shaded region has trapezoidal area of:

$$\begin{aligned} \text{area of trapezoid} &= \frac{\frac{1}{1000} + \frac{1}{1500}}{2} \cdot 250 \\ &= \frac{5}{24} \approx 0.2083 = 20.83\%. \end{aligned}$$

We can conclude that about 20.83% of days the oil field be expected to produce between 1750 and 2000 barrels.

Text Exercise 4.5.3

The probability distribution for gauging measurement uncertainties (error size when taking measurements of objects) is sometimes modeled by a trapezoidal-shaped distribution. Suppose the following graph represents such a distribution.

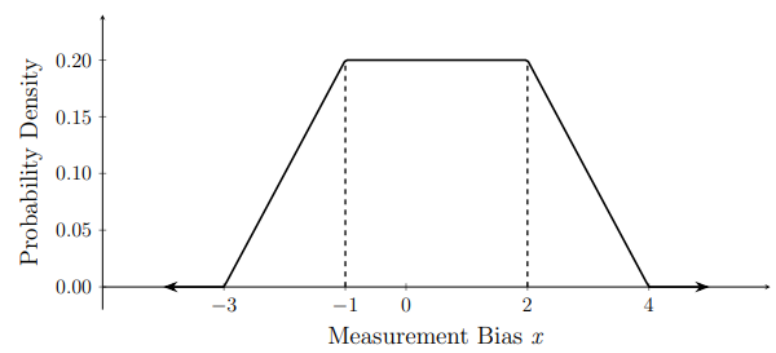


Figure 4.5.9: Probability distribution for measurement bias

1. Find the probability that a randomly selected value in this distribution is positive.

Answer

To find the probability that a randomly selected value is positive. we shade the area under the probability density curve on the interval from 0 to 4, and notice the region is a trapezoid.

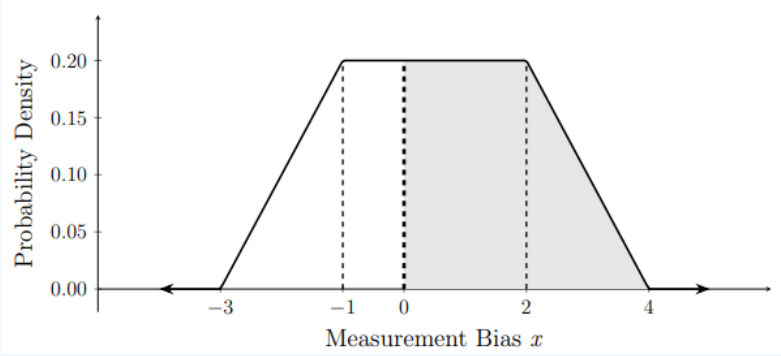


Figure 4.5.10 Finding $P(x > 0)$

The height of the trapezoid is 0.20, and the base lengths are 4 and 2. We thus find that the shaded area is $\frac{4+2}{2} \cdot 0.2 = 3 \cdot 0.20 = 0.60$. We thus have the probability that a randomly selected value is positive is 60%.

2. Find the probability that a randomly selected value in this distribution is at least 2; that is, find $P(x) \geq 2$.

Answer

To find the probability that a randomly selected value is at least 2, we shade the area under the probability density curve on the interval from 2 to 4, and notice the region is a triangle.

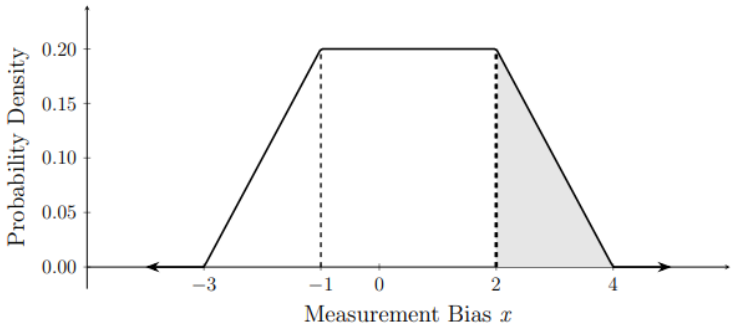


Figure 4.5.11: Finding $P(x \geq 2)$

The height of the triangle is 0.20, and the base is 2. We thus find that the shaded area is $\frac{1}{2} \cdot 2 \cdot 0.2 = 1 \cdot 0.20 = 0.20$. We thus have the probability that a randomly selected value is at least 2 is 20%.

3. Determine $P(-1 < x < 1.5)$.

Answer

To find $P(-1 < x < 1.5)$, we shade the area under the probability density curve on the interval from -1 to 1.5, and notice the region is a rectangle.

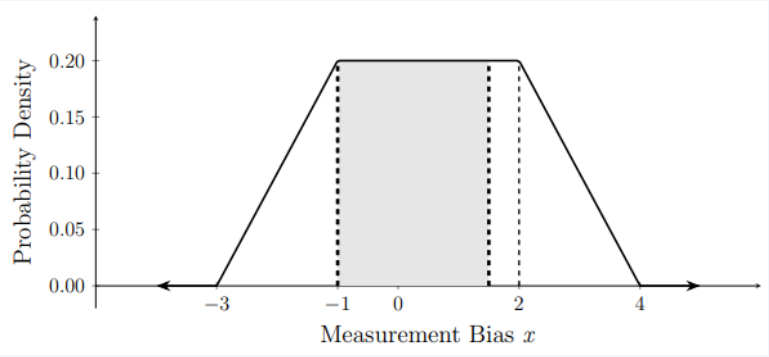


Figure 4.5.12 Finding $P(-1 < x < 1.5)$

The height of the rectangle is 0.20, and the base is $1.5 - (-1) = 2.5$. We thus find that the shaded area is $2.5 \cdot 0.20 = 0.50$. We thus have $P(-1 < x < 1.5)$ is 50%.

We continue to see how knowledge of geometric figures and creative geometric thinking on regions can help analyze the probability distributions of continuous random variables. Some distributions of continuous random variables have a more exciting and challenging distribution shape than those above. Let us examine some of those next.

Normal Distributions

One of the most commonly used probability distributions on continuous random variables is the normal distribution mentioned in Section 2.7. A **normal probability distribution**, also called the **Gaussian probability distribution**, is a bell-shaped, perfectly symmetric probability density curve that is centered above a mean value and has the specific property that the two changes of concavity on the density curve (called inflection points) occur at exactly one-standard deviation from the center mean location with the horizontal scale. As shown in Figure 4.5.13 below, a normal distribution is located with a horizontal scale solely by the knowledge of the mean μ value and the standard deviation σ.

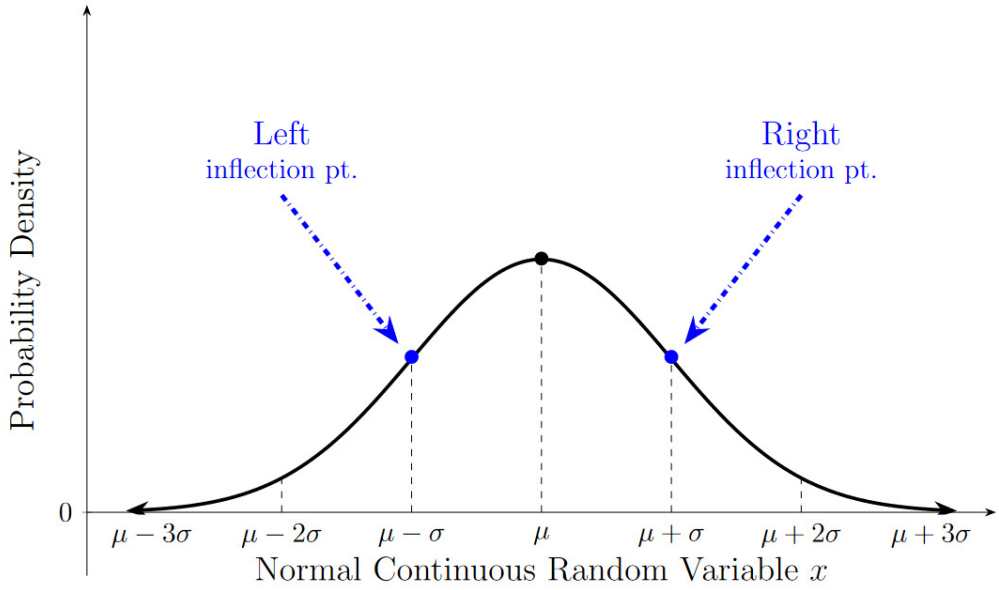


Figure 4.5.13: General Normal Distribution

As is true of all valid probability distributions on continuous variables, the total area under the curve is equal to $1 = 100\%$. Usually, we do not label the probability density axis (the vertical axis), but we always scale the horizontal axis with our continuous random variable of interest. The term "normal" is used because this distribution has surprised statisticians and others with how often it is found in the analysis of random events and as the shape in many continuous variable distributions from data-based histograms.

Naturally, as different mean and standard deviation values occur in the many normally distributed random variables, there is a whole family of distributions called "normal probability distributions." In Figure 4.5.14 below, we see four different normal distributions. We should notice how each normal distribution is controlled by its mean and standard deviation. The mean locates the center of the bell-shaped curve on a given horizontal axis scale, and the standard deviation controls the spread/width of the curve on the same scale. We should notice how the height of the bell-shaped curve is larger with smaller standard deviations and smaller with larger standard deviations. This should make sense to us as we must maintain an area of $1 = 100\%$ within the curve, so the curve's height must be associated with the spread. If sketching a distribution by hand, we will usually make the bell-shaped curve shape first, add a horizontal axis below the curve, and then scale that axis to meet the mean and standard deviation values required locations.

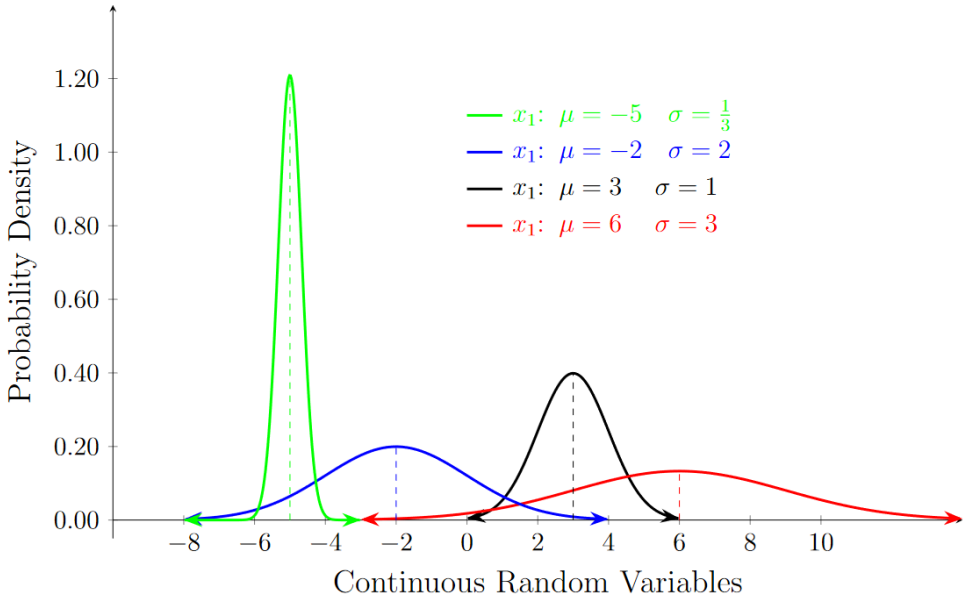


Figure 4.5.14: Four Different Normal Distributions

As mentioned in Section 2.7, there is one special normal probability distribution called the **standard normal** or **z-distribution**; this refers to a specific normal distribution that has $\mu = 0$ and $\sigma = 1$, producing the normal distribution curve shown in Figure 4.5.15 below.

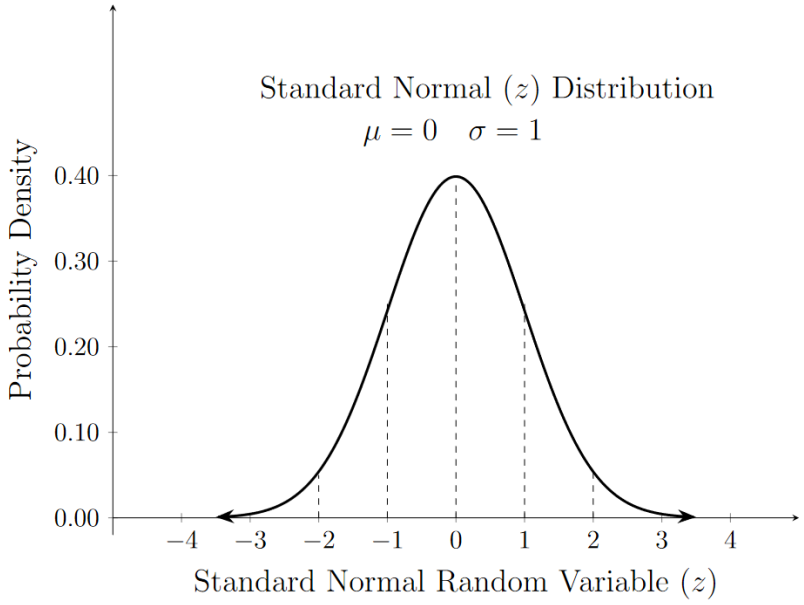


Figure 4.5.15: The standard normal distribution

All the various random variables x that are normal probability distributions can be converted to the standard normal distribution through use of the z-score or standardization computation of

$$z = \frac{x - \mu}{\sigma}$$

as discussed in Section 2.7. This is illustrated in Figure 4.5.16, in which a normal distribution with $\mu = 10$ and $\sigma = 2$ has its raw x-axis also rescaled to the standard normal distribution scale.

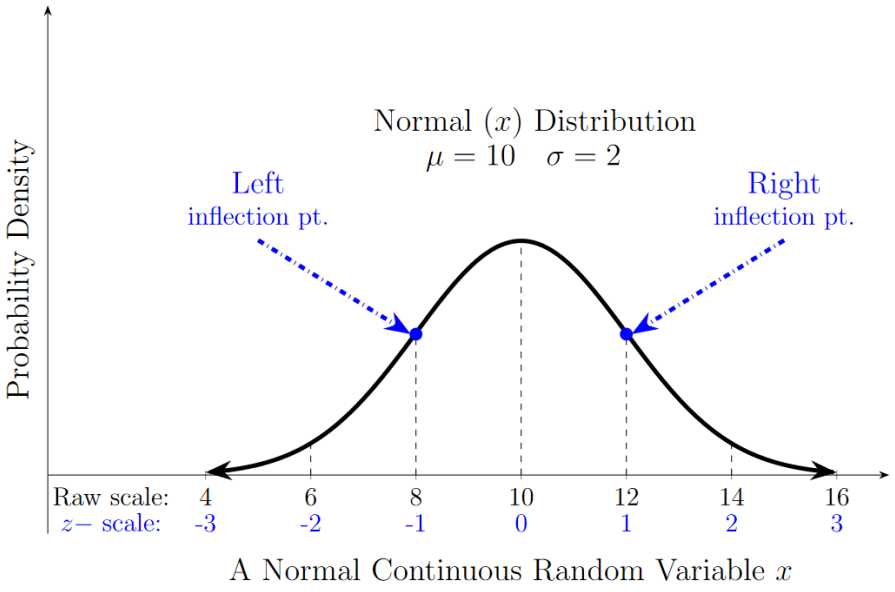


Figure 4.5.16: Standardized scaling on a non-standard normal distribution

We can convert from z-score to raw scale value in normal distributions by solving our equation for x :

$$x = \mu + z \cdot \sigma.$$

Let us now turn our focus on a quick review of the Empirical $68-95-99.7$ Rule, from Section 2.7. Our Empirical Rule gave us some approximate probability/area measures. As a reminder, we repeat the diagram in Figure 4.5.17below:

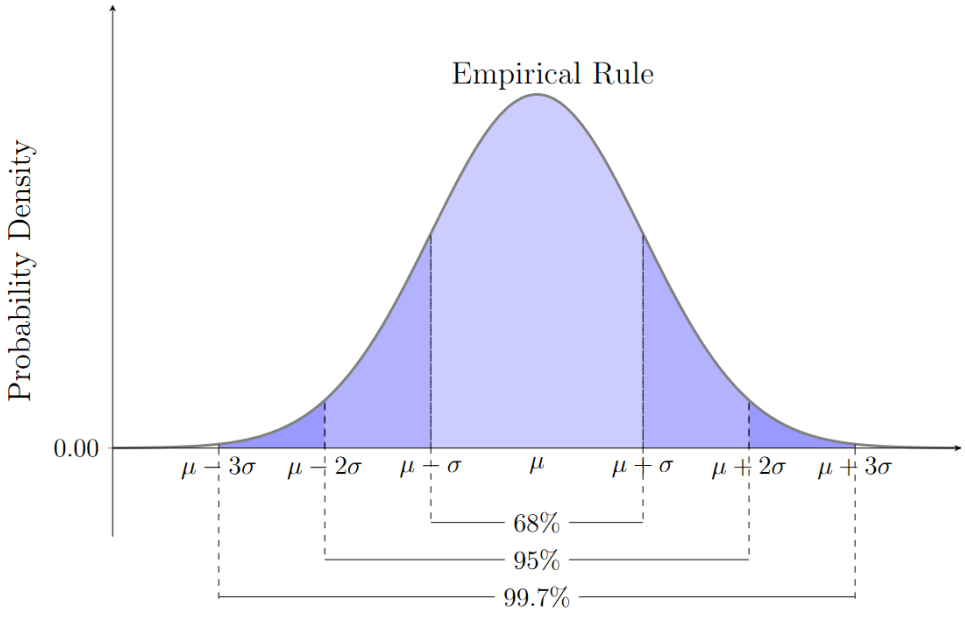


Figure 4.5.17: Empirical Rule on Normal Distribution

Text Exercise 2.7.7 had us working with this diagram and finding the percentage of observations (i.e., probability of occurrence) in the normal distributions. We will not repeat that collection of exercise questions here but leave any review, as necessary, to you. It is common to use these approximate area values when working with normal distributions, provided we are interested in values (outcomes) tied to the mean and standard deviations.

As the Empirical Rule was based on specific intervals tied to integer multiples of standard deviations from the mean, the rule is a bit limiting for other general interval choices in which we might be interested. For example, suppose the weights of thirty-year-old men in Chicago are normally distributed with $\mu = 190$ lbs. and $\sigma = 11.2$ lbs. By our Empirical Rule, the probability of randomly selecting a thirty-year-old man in Chicago weighing between $190 - 2 \cdot 11.2 = 167.6$ and $190 + 2 \cdot 11.2 = 212.4$ lbs. is approximately 95% (or stated equivalently, about 95% of thirty-year-old men in Chicago weigh between 167.6 and 212.4 lbs.) But what if we were interested in the interval of weights between 175 and 200 lbs, as shown in Figure $4.5.18$ below?

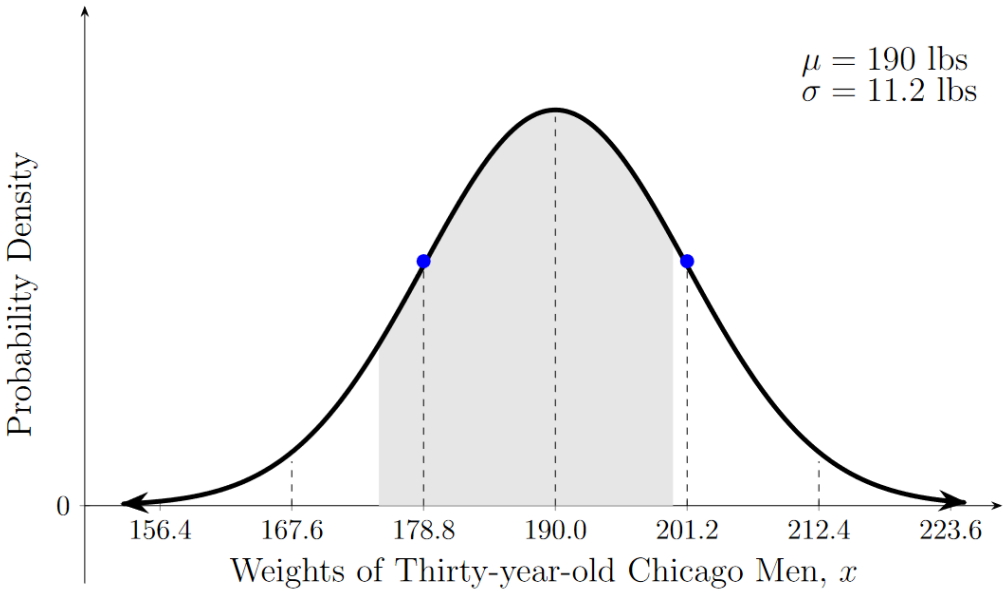


Figure 4.5.18: Normal distribution of the weights of thirty-year-old Chicago men

Our Empirical Rule does not apply to such varied intervals. However, if we could determine the area of this shaded region in this normal probability distribution, we could find the probability measure. This idea extends to any desired interval(s) we want to analyze. However, as the shaded region is not a basic geometric shape, as in our earlier work in this section, we cannot call on our knowledge of basic geometric formulas to find the area measures. We need other methods for handling such shaped regions. With this in mind, we will introduce technology-based methods in Section 4.6 for finding areas of any region(s) we want within any normal probability distribution. For now, we try some exercises to ensure we can sketch a described normal distribution or interpret key features of a given normal distribution graph.

? Text Exercise 4.5.4

1. Create graphs of normal distributions that meet the given descriptions. Include labeling of the horizontal axis with appropriate scaling (in standard deviated units) and axis titles:
   a. A soft drink bottler has data that suggest that the amount of drink actually placed in the cans by a specific bottling machine is normally distributed with $\mu = 12.1$ ounces and $\sigma = 0.5$ ounces (the machine is slightly over-filling on average from designed specifications.)
   b. The average consumption of electricity by electric four-door passenger vehicles is believed to be normally distributed with $\mu = 0.346$ kWh per mile with $\sigma = 0.022$ kWh per mile, where kWh stands for kilowatt hour.
   c. A tire company is about to begin large-scale manufacturing of a new tire made of newly developed materials. The tire's tread life has been tested, the research team found the tread life in miles produced a normal distribution with $\mu = 72,000$ miles and $\sigma = 7,000$ miles.

Answer

a. Given the key parameters of $\mu = 12.1$ ounces and $\sigma = 0.5$ ounces, we produce the following sketch, making sure to align our scale axis to this information by placing the value 12.1 directly below the peak of the normal distribution's PDF curve, and scaling out by values of 0.5 directly below the inflection points of our curve in order to incorporate the

key spread measure of the standard deviation. Then keeping this distance consistent, we scale out farther left and right with more standard deviated units on the axis.

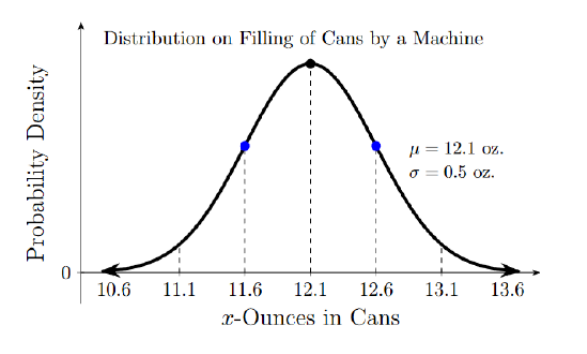


Figure 4.5.19 Normal distribution with $\mu = 12.1$ ounces and $\sigma = 0.5$ ounces

b. Given the key parameters of $\mu = 0.346$kWh per mile with $\sigma = 0.022$ kWh per mile, we produce the scaled normal distribution figure using the same approach as part a. directly above.

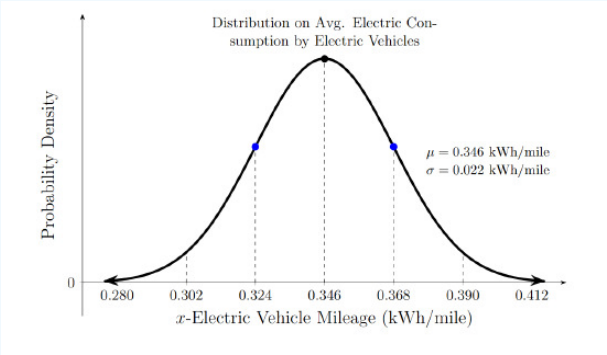


Figure 4.5.20 Normal distribution with $\mu = 0.346$ kWh and $\sigma = 0.022$ kWh

b. Since $\mu = 72,000$ miles and $\sigma = 7,000$ miles, we produce the following sketch:

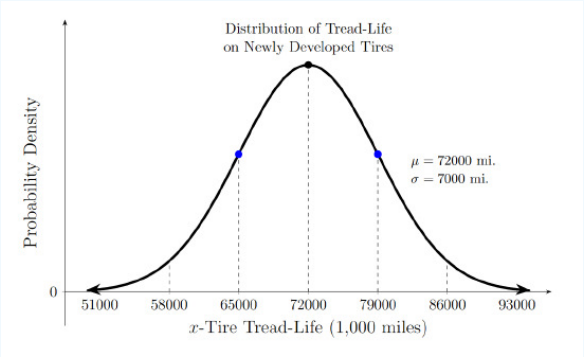
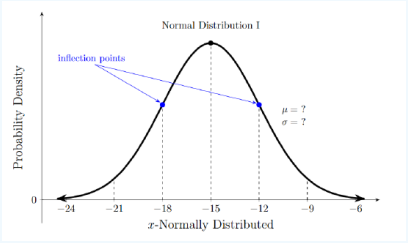
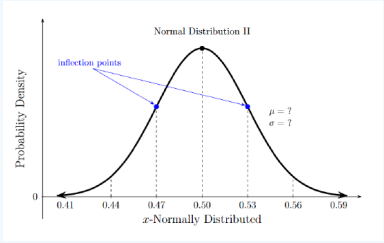


Figure 4.5.21 Normal distribution with $\mu = 72,000$ miles and $\sigma = 7,000$ miles

2. For each of these normal distributions, give the mean μ and the standard deviation σ of each graph.

I. 

II. 

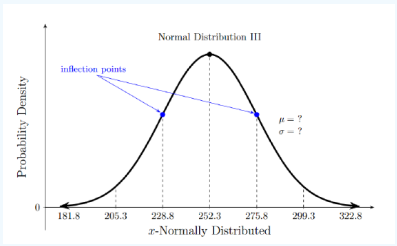
III. 

Figure 4.5.22: Normal distributions with various means and standard deviations

Answer

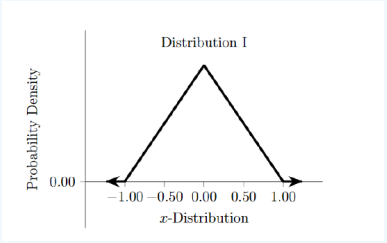
The location of the mean for each is the scale value directly below the highest point of the normal distribution. The standard deviation for each is the distance in the horizontal scale between the high point and either of the inflection points of the normal distribution. This produces the following results for each of the graphs.

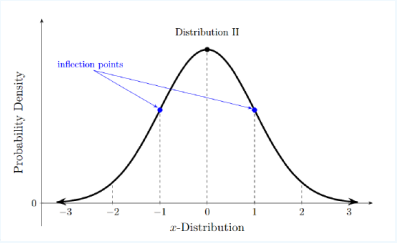
I. From the graphic of Distribution I, we see that $\mu = -15$ since the high point of the probability density curve is directly above that horizontal scale value. Also, since the labeled inflection points occur at a horizontal distance of $|(-15) - (-12)| = |(-15) - (-18)| = 3$, then $\sigma = 3$.

II. From the normal distribution figure, $\mu = 0.50$ since the high point of the probability density curve is directly above that scale value. Also, since the labeled inflection points occur at a horizontal distance of $|0.50 - 0.53| = |0.50 - 0.47| = 0.03$, then $\sigma = 0.03$.

III. From the given normal distribution, $\mu = 252.3$ since the high point of the probability density curve is directly above that value. Also, since the labeled inflection points occur at a horizontal distance of $|252.3 - 275.8| = |252.3 - 228.8| = 23.5$, then $\sigma = 23.5$.

3. Which of these is the standard normal distribution and which are not.

I. 

II. 

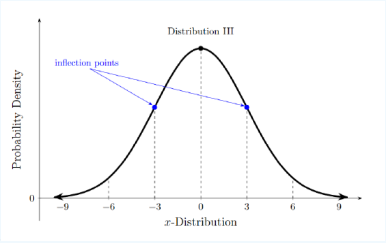
III. 

Figure 4.5.23: Various probability distributions

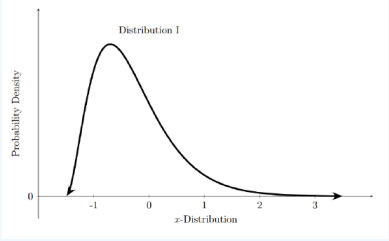
Answer

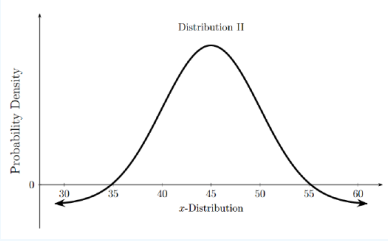
Distribution II is the standard normal distribution as the function is a bell-shaped symmetrical distribution with high point located above the horizontal axis value of 0 (implying a $\mu = 0$) and inflection points at 1 unit away in terms of the horizontal axis scale (implying a standard deviation of $\sigma = 1$.

Distribution I is a symmetrical distribution about 0, however, the shape is triangular and not bell-shaped. So, Distribution I is not a normal distribution.

Distribution III is a bell-shaped symmetrical distribution with high point located above the horizontal axis value of 0; however the inflection points at 3 units away on the horizontal axis scale are implying a standard deviation of $\sigma = 3$. So although a normal distribution, the Distribution III is definitely not the standard normal distribution.

4. Explain why each of the graphs below are not representative of a normal distribution.

I. 

II. 

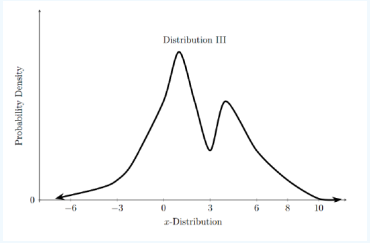
III. 

Figure 4.5.24: Various graphs to be considered

Answer

Distribution I is not a symmetrical distribution (though a bit bell-shaped.) Most would consider Distribution I a skewed right distribution, so not a normal distribution.

Distribution II is a nice symmetrical bell-shape, but is not a probability density function since some of the function values are below the horizontal axis. PDF values can only be non-negative. So Distribution II is not a normal distribution as all normal distributions are represented by valid PDFs.

Distribution III is neither symmetrical or bell-shaped, and hence not a normal distribution.

As we continue through the course, we will often find ourselves working with normal probability distributions to answer questions. But there are a few other distribution curves that we will find ourselves working with as well. We briefly examine two more such families of distributions next.

Other Distribution Families

Although normal distributions are arguably the most frequently examined and applied distribution in introductory statistics, we examine other families of probability distributions with important statistical applications. First, we will briefly discuss the **Student's t-distributions.**

These probability distributions are sometimes called $t-$distributions. As shown in Figure 4.5.25,these distributions initially appear to be much like the family of normal distributions since they are also a symmetric bell-shaped family of distributions and the total area under the density curves is $1 = 100\%$.

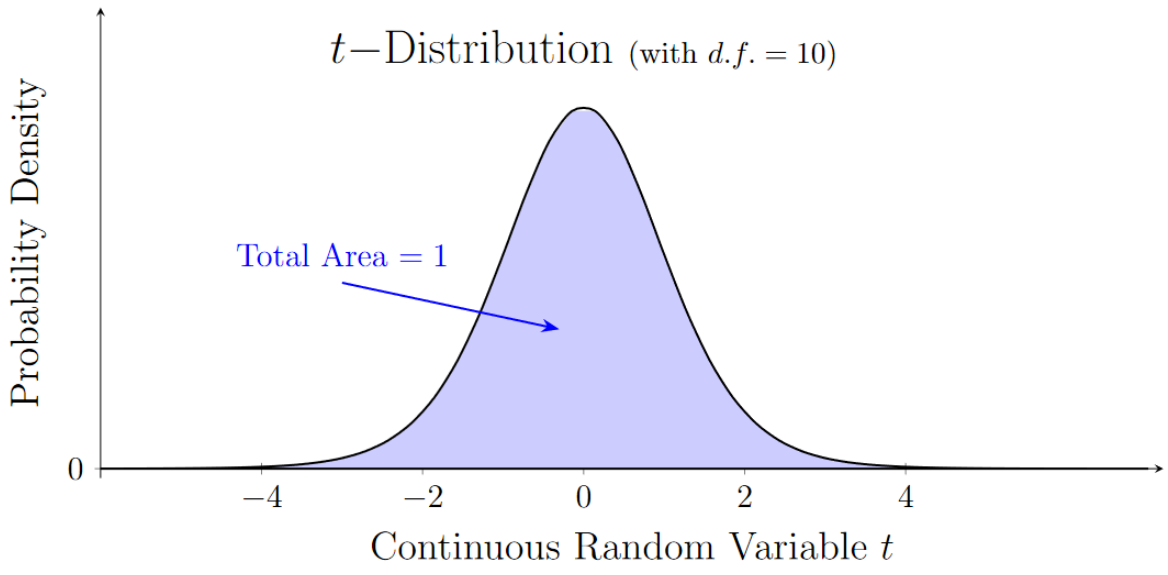


Figure 4.5.25: A Student's t-distribution with $d.f. = 10$

All t-distributions are symmetric with a mean of $\mu = 0$, but their inflection points do not occur at one standard deviation from the mean as in normal distributions. The t-distributions have thicker tails than the standard normal distribution; meaning, a t-distribution is more likely to see an outcome far from the mean than a normal distribution. While normal distributions are defined by a given μ and σ, the spread of the t-distributions is controlled by a value called a degree-of-freedom, $d.f.$. In future sections, this value will be related to sample size, and the origin of the name will be clearer. As this degree of freedom value increases in size, the related t-distributions become more and more like the standard normal distribution. Figure 4.5.26 shows sketches of several t-distributions as well as the standard normal distribution for comparison.

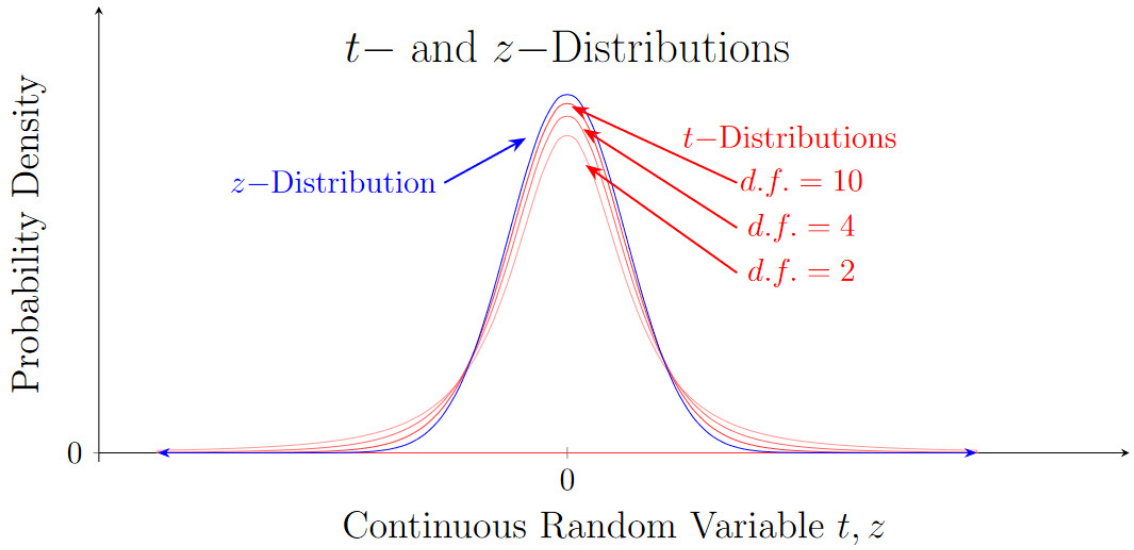


Figure 4.5.26: Several $t-$distributions and the standard normal distribution

These t-distributions will become very important in our future work. For now, we understand that they are another particular family of probability distributions and that probability in the distribution can be determined by finding area measures of specified regions.

Optional t-distribution discussion for the mathematically inclined

The behavior of t-distributions can be explained by the fact that they are defined using a rational function; whereas, normal curves are defined using an exponential function. For example, the formula for the probability density function of a t-distribution with 1 degree of freedom is given by

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

In general, if one has d degrees of freedom, the probability density function is given by

$$f(x) = \frac{c_d}{(1+x^2/d)^{(d+1)/2}},$$

where c_d is a suitable constant which makes the total area $= 1$.

A third family of probability distributions common to beginning statistics analysis are the Chi-squared distributions (written in the Greek letter, χ^2-distributions). This Greek letter χ is pronounced as $k\bar{i}$, similar to the pronunciation of the first two letters in the word *kite*. This family of distributions is not symmetrical but instead positively skewed in shape with non-zero density values and all domain values larger than 0; see Figure 4.5.27below.

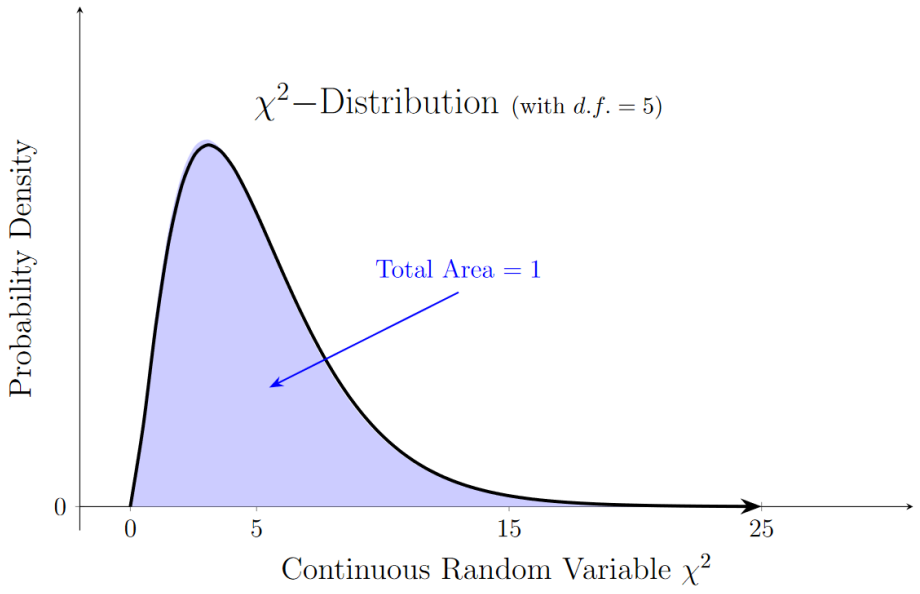


Figure 4.5.27: A χ^2-distribution with $d.f. = 5$

Similar to the t-distributions, χ^2-distributions are controlled by a degree of freedom value. In Figure 4.5.28 are sketches of several χ^2-distributions. Notice in the figure that as the degree of freedom value increases in size, the distribution approaches a bell-shaped curve. Some interesting properties of these distributions are that the high point occurs two units in the scale before the degree of freedom value and that the expected value is its degree of freedom value.

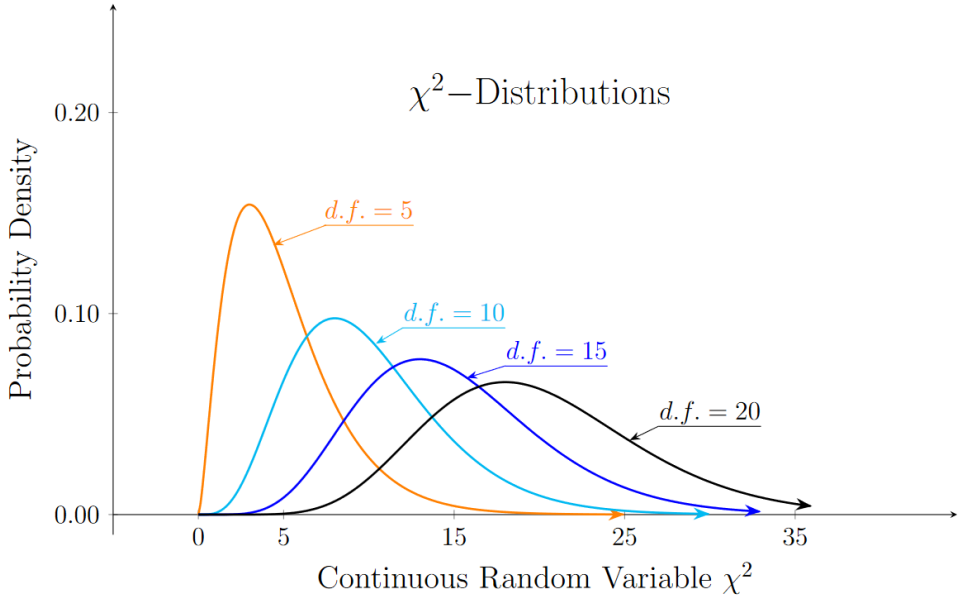
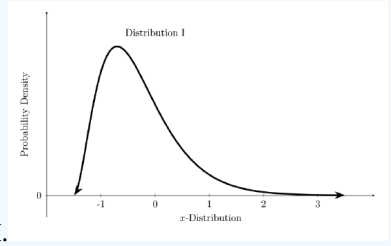


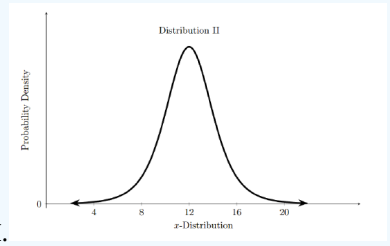
Figure 4.5.28: Several χ^2-distributions

As with the $t-$distributions, we will not formally delve into the χ^2 density functions mathematical formulas. It will be sufficient for us to understand the shape of these distributions and, with future work, recognize when they are to be used.

Text Exercise 4.5.5

1. Which of the following cannot possibly be a t-distribution? Explain.

I. 

II. 

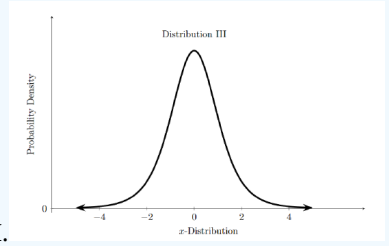
III. 

Figure 4.5.29: Various probability distributions

Answer

Distribution I cannot be a t-distribution since it is a skewed distribution. All t-distributions are bell-shaped and symmetric about the horizontal scale value of 0.

Distribution II, although it is bell-shaped and symmetrical, cannot be a t-distribution since the curve is symmetric about horizontal scale value of 12 instead of 0.

Distribution III might be a t-distribution since the curve is bell-shaped and symmetric about the horizontal scale value of 0. We do note that there are other probability density curves that might make a very similar shape and be positioned as shown in a scale axis. To know for sure, more information would be needed, such as several probability density values.

2. The graph below gives three t-distributions and the standard normal distribution. Which of the t-distributions has the largest degree-of-freedom value? Explain.

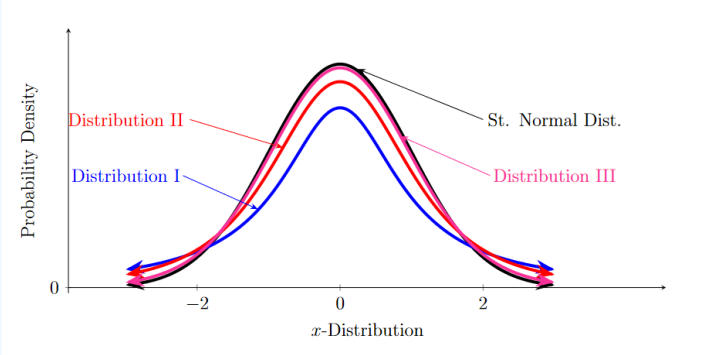
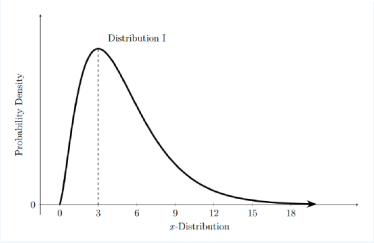


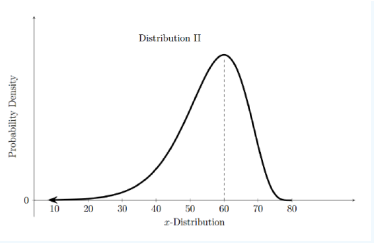
Figure 4.5.30: Various t-distributions plotted with the standard normal distribution

Answer

As the degree-of-freedom increases on t-distributions, the distributions begin to come very close to the standard normal distribution. Hence Distribution III must have the largest degree-of-freedom value. We also note that Distribution I must have the smallest degree-of-freedom value since the shape of the t-distribution is wider and shorter than the rest of the distributions shown.

3. Which of the following are possible χ^2-distributions? Estimate the degree of freedom value for any that appear to be possible χ^2-distributions.

I. 

II. 

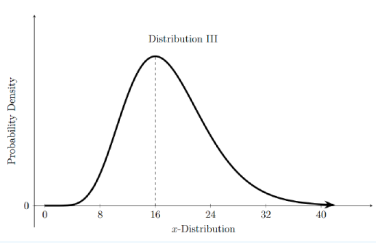
III. 

Figure 4.5.31: Various probability distributions

Answer

Distribution I can possibly be a χ^2-distribution since the graph is a positively skewed distribution with only density measures tied to non-negative horizontal scale measures. Because the high-point on the curve occurs at a horizontal scale value of 3, the degree-of-freedom value is $3+2\ =5$.

Distribution II cannot be a χ^2-distribution since the graph is skewed negatively instead of positively.

Distribution III can possibly be a χ^2-distribution since the graph is a positively skewed distribution with only density measures tied to non-negative horizontal scale measures. Because the high-point on the curve occurs at a horizontal scale value of 16, the degree-of-freedom value is $16+2\ =18$.

4. What would be the most likely shape (uniform, normal, skewed right similar to χ^2 for each of the random variables described below?

   a. Ages of coins in circulation

   b. Birth weight of babies born in Hays during the time interval of $2020-2023$.

   c. Position of one tire valve (in degrees) on vehicle wheel when the vehicle stops at various times in the day

   d. IQ scores of all senior class students in the United States

   e. Income of adult Kansas residents

Answer

a. Ages of coins in circulation would likely be a positively skewed distribution since there are many more coins of young ages, and very few coin of older ages.
b. Birth weight of babies born in Hays during those years is likely to be a normal distribution, there will be a few light babies and a few heavy babies born, but most babies will be around the same weight as the average.
c. The position of the tire valve on a vehicle wheel (as measured in degrees) is likely to be uniformly distributed. One position is just as likely as another in some random-length trip with the vehicle.
d. IQ scores are likely to be normally distributed, there will be a few high IQ individuals and a few low IQ individuals, but most senior class students in Kansas will have close to the average IQ.
e. Incomes are likely to be a positively skewed distribution. Incomes cannot be negative (in a normal meaning of income), and incomes will increase to some high point, before trailing off to those few making very high incomes.

There are many other families of distributions in statistics; however, our current list will be sufficient for an introductory course. We now return to the issue of finding area measures of regions in all these various probability distributions on continuous random variables.

Geometric Connections on Related Regions in Continuous Distributions

Before moving to a new section, we will explore how we can use basic geometric reasoning with the area of regions in all types of continuous random variable distributions to develop a general approach that works for various families of continuous probability distributions. In much of our future work with these various distributions, we need to tie any interval-designated region of interest to a left-region: regions that cover the left tail of our distribution. We must be discussing intervals that have the less than inequality in them: either $x<a$ or $x\leq a$ where a is an outcome value for the random variable.

We will use what is given if our region of interest is already a left region. For example, suppose we are working with a specific t-distribution in which we need to find $P(t < 1.2)$. Creating a sketch of our region as shown in Figure $4.5.32$, we notice that our region completely covers the left tail of the distribution. In the next section, we will see how our computational technology can produce the area measure of only left-tail regions.

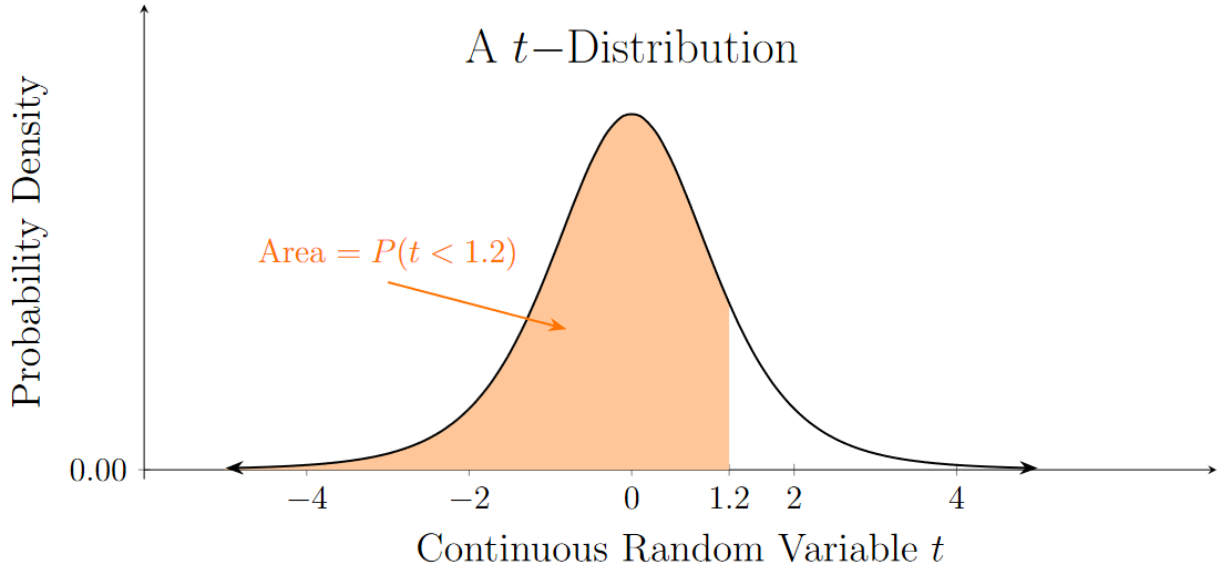


Figure $4.5.32$: Left-tailed region in a probability distribution

But what if our region of interest is a right-tail region? For example, suppose we are working with some normal probability distribution in which we desire to find $P(x \geq 25)$. Creating a sketch of our region as shown in Figure $4.5.33$, we notice that our region covers the right tail of the distribution and not the left.

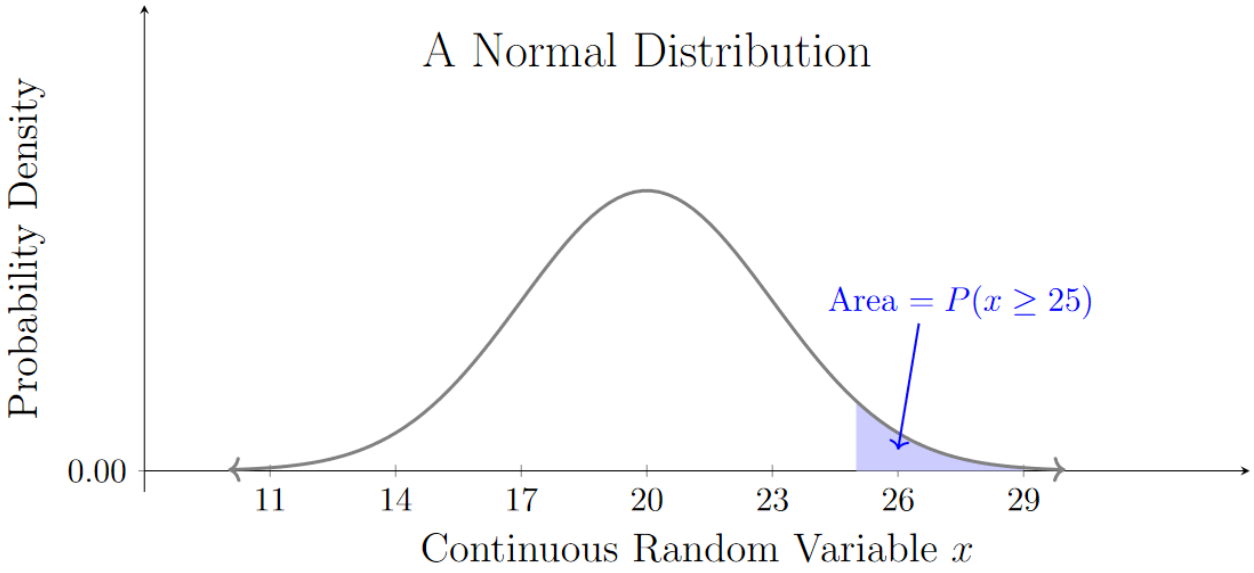


Figure $4.5.33$: Right-tailed region in a probability distribution

Using our geometric reasoning and complement property on probabilities, we notice that the white region under the curve is the complement to the shaded region. The complement probability $P(x < 25)$ is a left-tailed region. We can know $P(x \geq 25)$ by

relating to $1.0000 - P(x < 25)$. Using this geometric reasoning, we can relate any right-tailed region to a left-tail region by applying our complement concept.

Some regions are neither left- or right-tailed regions. For example, suppose we are in a specific χ^2-distribution and we need to find the probability value $P(4 \leq \chi^2 \leq 8)$. Creating a sketch of our region in Figure 4.5.34,we see that our region covers neither a left- or a right-tailed region. We have what is called a central or between region.

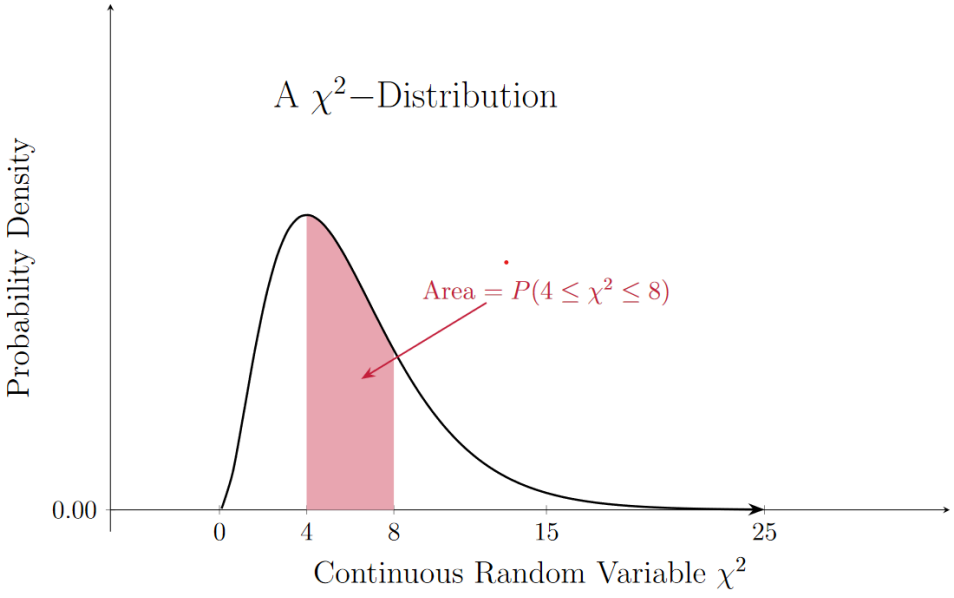


Figure 4.5.34: Between/Central region in a probability distribution

Using basic geometric reasoning, we can relate this region of interest to the left-tailed areas. Notice that the left region associated with the interval inequality $\chi^2 \leq 8)$ covers our region of interest but also the undesired left-tailed region related to the inequality $\chi^2 \leq 4$. Suppose we *remove* the left region associated to $\chi^2 < 4$ from the larger left region related to $\chi^2 \leq 8$, then all that remains is the original region of interest between 4 and 8. See this illustrated in the related diagram below.

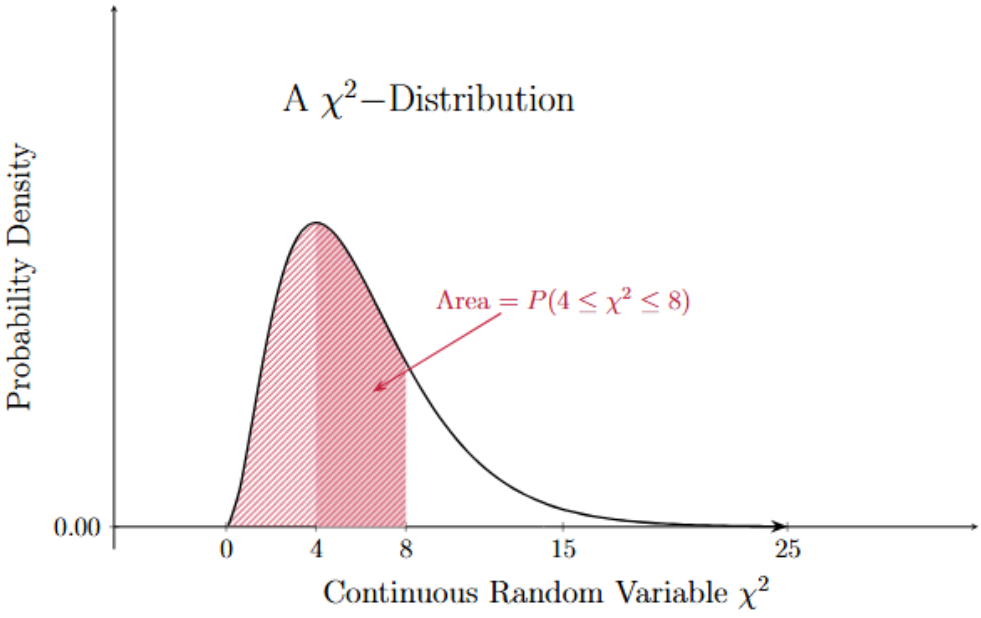


Figure 4.5.35: Visualization of a central region as the difference of two left regions

In terms of probability notation, we have

$$P(4 \le \chi^2 \le 6) = P(\chi^2 \le 6) - P(\chi^2 \le 4).$$

All of the above is summarized in Figure 4.5.36 (we note that although the figure contains only bell-shaped distributions, the shape of the specific continuous probability distribution does not change the general geometric reasoning).

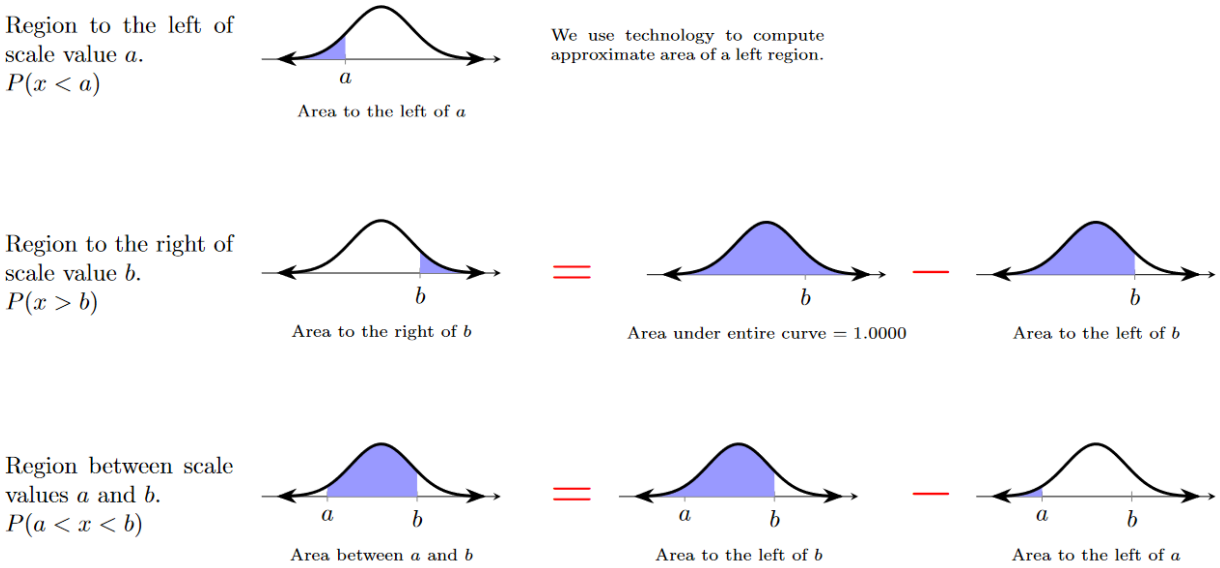


Figure 4.5.36: Transforming any region into a related left-tail region

For any interval of interest, a right-region area can be found using a complement concept: the total area of 1 minus the related complement left-region area. A between-region area can be found by the total left-region area from the right boundary of the interval minus the total left-region area from the left boundary of the interval. Stated in symbolic representation:

$$P(x > b) = 1.0000 - P(x \le b)$$
$$P(a < x < b) = P(x < b) - P(x \le a).$$

Once we fully grasp how to relate any region of interest in a continuous variable's probability distribution to only left-region measures, we should be able to reason similarly if only right-region area measures are available. In this chapter's next section, we will introduce technology that computes only left-region area measures.

? Exercise 4.5.6

Describe how the area of the shaded region in each of the given probability distributions can be expressed in terms of left-tail region(s). If the shaded region is already of left-tailed type, state so.

1. $P(x \ge 12)$ in the distribution

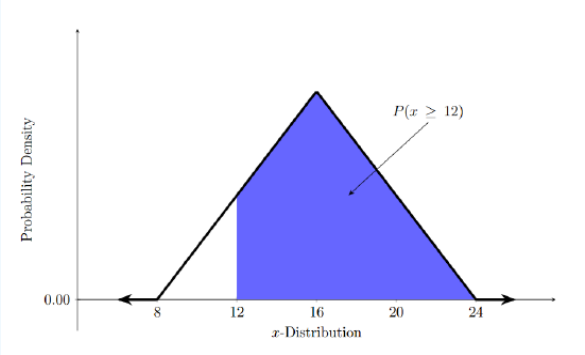


Figure 4.5.37: Probability distribution with shaded region

Answer

Since this is a right-tailed region, we use the complement approach with a left-tailed region:

$$P(x \geq 12) = 1.0000 - P(x < 12).$$

2. $P(x < 0.45)$ in the distribution

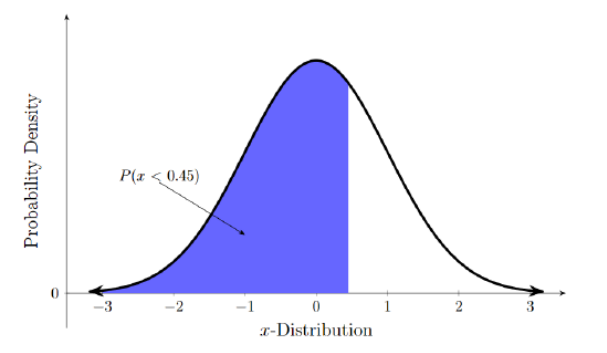


Figure 4.5.38: Probability distribution with shaded region

Answer

Since this is already a left-tailed region, no change is needed:

$$P(x < 7) = P(x < 7).$$

3. $P(-3 < x \leq 6)$ in the distribution

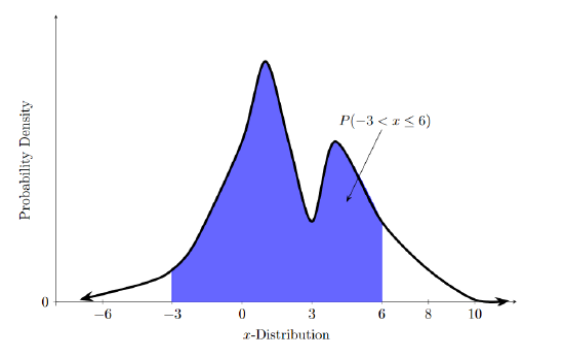


Figure 4.5.39: Probability distribution with shaded region

Answer

Since this is a central region, we use subtraction between two left-tailed regions:

$$P(-3 < x \leq 6) = P(x < 6) - P(x \leq -3).$$

4. $P(x > 16)$ in the distribution

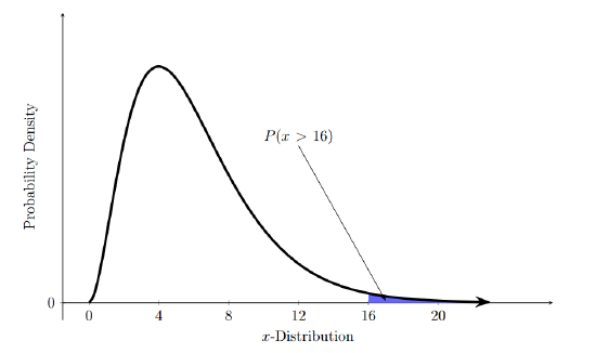


Figure 4.5.40: Probability distribution with shaded region

Answer

Since this is a right-tailed region, we use the complement approach:

$$P(x > 16) = 1.0000 - P(x \leq 16).$$

5. $P(6 < x < 16)$ in the distribution

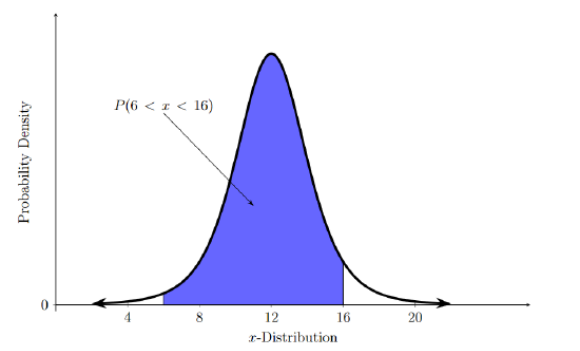


Figure 4.5.41: Probability distribution with shaded region

Answer

Since this is a central region, we use subtraction between two left-tailed regions:

$$P(-8 < x \leq 10) = P(x < 10) - P(x \leq -8).$$

Summary

This section introduced several different families of continuous random variable probability distributions. We used geometric area formulas to determine the probability of outcomes for some random variables. We also looked into other notable families of continuous random variable probability distributions, such as the normal and t-distributions. Finally, we examined geometric region relationships on these distributions and how any interval region can be considered within only left-tail region(s).

In the next section, we focus on using these area relationships with technology-based cumulative distribution functions in normal distributions. These special area accumulation functions will provide accurate area measures of regions in these distributions.

4.6: Accumulation Functions And Area Measures in Normal Distributions

Learning Objectives

- Define an accumulation function for continuous probability distributions
- Use an accumulation function for the standard normal (z-)distribution to find area measures of regions in the standard normal distribution
- Use the inverse of an accumulation function for the standard normal (z-)distribution to find the location for the specified region's area measures
- Standardize non-standard normal distributions to find area measures and scale locations
- Use an accumulation function for general normal distributions to find area measures of regions in those distributions
- Use the inverse of an accumulation function for general normal distributions to find scale location for specified region's area measures
- Use spreadsheet functions of NORM.S.DIST, NORM.S.INV, NORM.DIST, and NORM.INV appropriately for finding needed values in normal distributions

Review and Preview

We have discussed the relationship between the area of regions within a continuous random variable's probability distribution and the probability of occurrence in relation to that variable. We also examined several families of distributions. Lastly, we noted how any region of interest in these distributions could always be related to left-regions. We now focus on how to produce these left-region area measures on normal distributions using technology. Once we reasonably master these concepts in relation to normal distributions, similar ideas are used in t-distributions and χ^2-distributions, as well as many other specialized distributions.

Accumulation Functions of Area

We have discussed the importance of the determining the area of regions within probability distributions since the probability of selecting an outcome in the region formed by an interval is equal to its area. It became more difficult to determine the area if the regions of interest were not basic geometric shapes. Specifically, if our regions were rectangular, we could easily compute the area of such regions.

In general, the area of a region in our distributions can be sliced up into thin slices. We can approximate the area of each of these thin slices by summing the areas of rectangles that are close in height to the thin slices. This is illustrated in Figure 4.6.1 below.

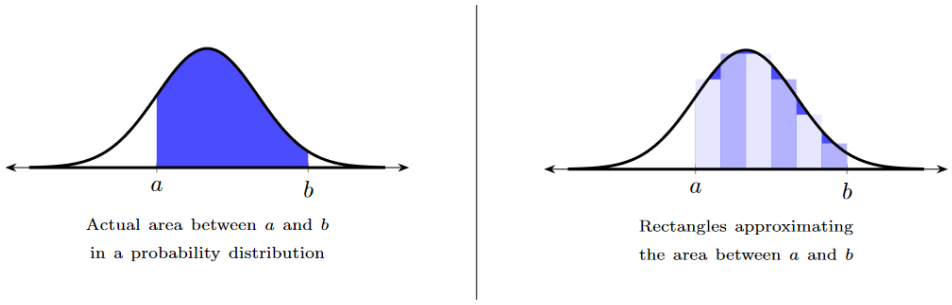


Figure 4.6.1: Probability distribution region being approximated by thin rectangles

Although it is messy work for humans to complete this process with only ten or twenty slices, computing technology efficiently calculates with the use of hundreds or even thousands of thinner and thinner slices on the region. As the number of slices for a fixed region gets larger, the approximating rectangles get thinner, and the approximation from the sum of areas on the thin rectangles gets closer to the actual area of the original region. At one time, large tables of area values were produced to list the approximation sums. Now this process has been programmed for several probability distributions, producing specialized **accumulation functions (also called cumulative distribution functions)** that provide highly accurate approximations to the area of regions in most common probability distributions. Different statistical software might name and program their accumulation functions differently; we will focus on the accumulation functions within spreadsheets in our following work. We start with an accumulation function for finding accurate left-tail regions in the standard normal probability distribution.

Area Measures for the Standard Normal Distribution

We begin with an **accumulation function for the standard normal distribution**. The name and syntax of this function can vary depending on the technology being used. The name of the accumulation function in Excel is NORM.S.DIST. We note that the spreadsheet function name corresponds to the distribution we are discussing. This function requires we provide it with a specific z-score value; the function will then return the area of the region to the left of that z-value when choosing the TRUE option to accumulate left area.

Due to the symmetry of the standard normal curve, we know that $P(z < 0) = 0.5000 = 50\%$ as shown in Figure 4.6.2. If we enter $= \text{NORM.S.DIST}(0, \text{TRUE})$ as a function in a spreadsheet cell, the spreadsheet returns the left-area measure of 0.5000 with appropriate cell formatting as shown in the spreadsheet image.

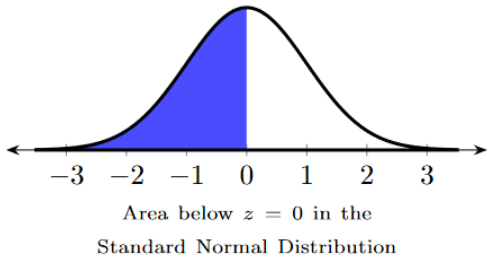


Figure 4.6.2: Standard normal distribution with shaded area

As a shortcut, we can enter the digit 1 instead of typing out the word TRUE when using this accumulation function. If we use FALSE, the function returns only the height measure of the density function but not an area measure. We will almost always want to use this function for area accumulation, but we must remember the function only returns left-tail area measures. If we find any other region, we must adjust our computation work as discussed in Section 4.5. In general, the syntax of this accumulation function is $= \text{NORM.S.DIST}(z\text{-score}, \text{TRUE})$ or the slightly shorter version of $= \text{NORM.S.DIST}(z\text{-score}, \mathbf{1})$.

Suppose we wish to find $P(z \leq -2)$, another left-tail region as shown by our graph of the standard normal distribution below. We found an area measure approximated at $0.0250 = 2.5\%$ in Section 2.7 through the use of the Empirical $(68 - 95 - 99.5)$ Rule.

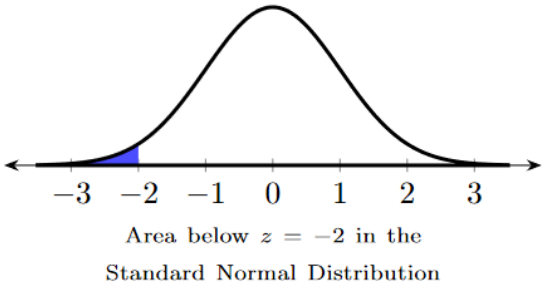


Figure 4.6.3: Standard normal distribution with shaded area

We can also use our spreadsheet accumulation function to find this area:

$$\begin{aligned} P(z \leq -2) &= \text{NORM.S.DIST}(-2, \text{TRUE}) \\ &= \text{NORM.S.DIST}(-2, 1) \\ &\approx 0.02275 = 2.275\%. \end{aligned}$$

For a random variable that possesses the standard normal distribution, we consider it unusual (since $2.275\% \leq 5.000\%$) to have a z-score that is at most -2. We note that this value of 2.275% from our technology's accumulation function is more accurate than the estimate from the Empirical Rule. Generally, we will use our technology to generate more accurate measures instead of the less accurate values computed by the Empirical Rule.

Suppose we want to find $P(z > 1.25)$, a right-tail region in the standard normal distribution as shown below. We must turn this into a left-tail region calculation to use our accumulation function.

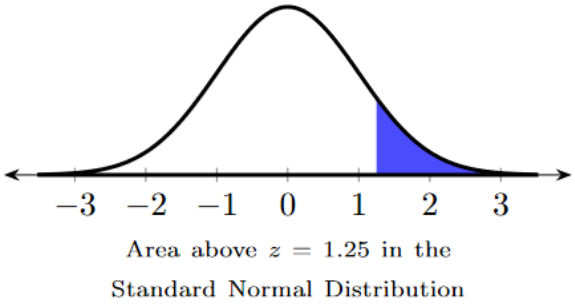


Figure 4.6.4: Standard normal distribution with shaded area

We should recognize the need to use the complement to relate to left-tail regions, producing:

$$\begin{aligned} P(z > 1.25) &= 1 - P(z \leq 1.25) \\ &= 1 - \text{NORM.S.DIST}(1.25, 1) \\ &\approx 1 - 0.89435 \\ &= 0.0.10565 = 10.565\%. \end{aligned}$$

Since it can be easy to make entry errors when using our spreadsheet functions, we compare our value to the region shaded in the graph. The shaded region does seem to be a small portion of the entire distribution, and our resulting value of 10.565% appears aligned to our graph. Until we have a strong mastery of the ideas, we will sketch the graph of the region of interest to aid us in proper accumulation function use and to verify the reasonableness of our computed area. Once mastery of these ideas is achieved, we encourage mental visualization of the graph of the distribution and showing work with the accumulation function.

? Text Exercise 4.6.1

Sketch graphs and determine the solutions of the following probability problems.

1. Find $P(z \geq -2)$.

Answer

After producing our graph representing $P(z \geq -2)$ (shown below), we notice we are working in a right-tailed region, but recall that the NORM.S.DIST function is for producing left-tailed area measures only.

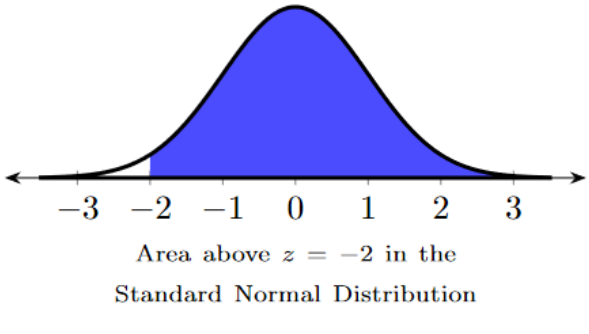


Figure 4.6.5: Standard normal distribution with shaded area

We must adjust the computation with the complement to find the right-tailed area in this situation.

$$\begin{aligned} P(z \geq -2) &= 1 - P(z < -2) \\ &= 1 - \text{NORM.S.DIST}(-2, 1) \\ &\approx 1 - 0.0228 \\ &= 0.9772 = 97.72\% \end{aligned}$$

Remembering our quick check, we notice that the size of the shaded region in the graph seems to align with this proportional measure of 97.72%. Thus, 97.72% of the standard normal distribution's area is to the right of -2. Or equivalently, there is a 97.72%probability of randomly selecting a z-score outcome that is at least -2 in value.

2. Find $P(-2 \leq z \leq 1)$.

Answer

After producing our graph (shown below), we notice we are working in a "between" region. In this situation, we again must make a computational adjustment with a difference calculation.

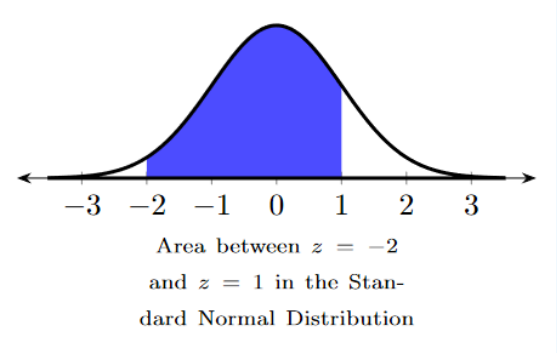


Figure 4.6.6 Standard normal distribution with shaded area

We subtract two left-tail areas to find the desired region's area measure.

$$\begin{aligned} P(-2 \leq z \leq 1) &= P(z \leq 1) - P(z < -2) \\ &= \text{NORM.S.DIST}(1, 1) - \text{NORM.S.DIST}(-2, 1) \\ &\approx 0.8413 - 0.0228 \\ &= 0.8185 = 81.85\% \end{aligned}$$

There is an 81.85%probability of randomly selecting a z-score that is between -2 and 1 in value.

3. Find $P(z \leq -1.25)$.

Answer

After producing our graph (shown below), we notice we are working in a left-tail region. No computational adjustment is needed to use our accumulation function in this situation.

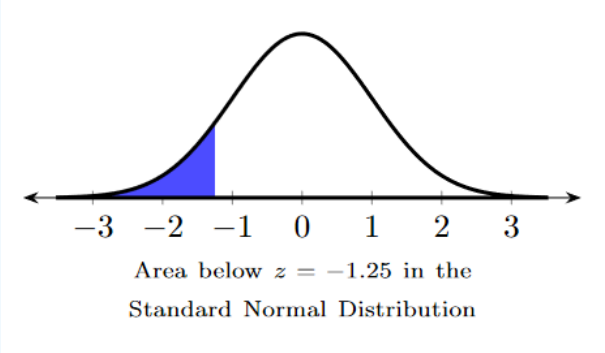


Figure 4.6.7 Standard normal distribution with shaded area

To find the area, we compute it with our spreadsheet function.

$$\begin{aligned} P(z \leq -1.25) &= \text{NORM.S.DIST}(-1.25, 1) \\ &\approx 0.1056 = 10.56\% \end{aligned}$$

After a quick check, we can confidently say that $10.56\%$of the standard normal distribution's area is to the left of -1.25. We also note that such an outcome in the random variable is not unusual since the probability measure is more than 5%.

4. Find the proportion of the standard normal curve between -1.5 and 0.5.

Answer

Based on our graph (shown below) of the given information, we notice we are again working in a "between" region. We must make the computational adjustment using another difference calculation.

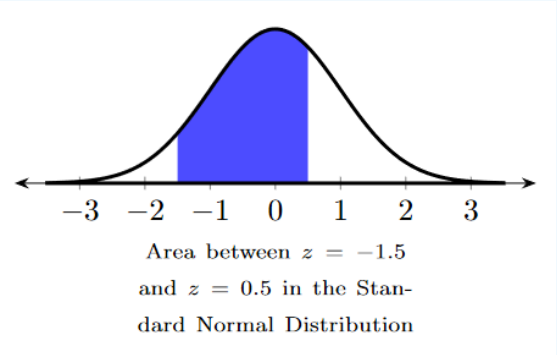


Figure 4.6.8 Standard normal distribution with shaded area

To find the area, we subtract the two left-tail areas.

$$\begin{aligned} P(-1.5 < z < 0.5) &= P(z < 0.5) - P(z \leq -1.5) \\ &= \text{NORM.S.DIST}(0.5, 1) - \text{NORM.S.DIST}(-1.5, 1) \\ &\approx 0.6915 - 0.0668 \\ &= 0.6247 = 62.47\% \end{aligned}$$

There is a 62.47%probability of randomly selecting a z-score outcome that is between -1.5 and 0.5 in value. Notice that this is the proportion of outcomes in this interval.

Working with decimal or fractional valued z-scores requires no adjustment in our thinking or work. The "messiness" of the numbers involved or produced should not impact our established reasoning or computational work.

Now, there will also be occasions in which we need to reverse the process above; that is, given the description of a region and its area measure, what is/are the z-score(s) that produce that region? For example, suppose we wish to know the one z-score that separates the lower 5% region of the standard normal distribution from the upper 95% region. Stated another way: what is the 5^{th} percentile of the standard normal distribution? This is illustrated in Figure $4.6.9$ below:

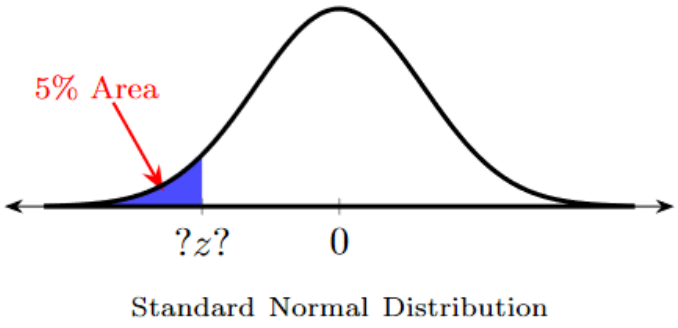


Figure 4.6.9: A region of the standard normal distribution with unknown $z-$score

With experiences from above, one might just guess-and-test reasonable z-score values with our NORM.S.DIST function to find an approximate value, a, for which $P(z < a) = 5\%$. This approach may take some time. Instead, we are blessed with the mathematics of inverse functions. We have a spreadsheet function called NORM.S.INV. Given any left-tail region's area, this function will compute the associated right boundary z-score that forms that region, an inverse process to the accumulation function. This function has the syntax $=$ NORM.S.INV(left-tail area measure between **0** and **1**). Since our left-tail area is $5\% = 0.05$, we compute in our spreadsheet to produce these results.

$$\begin{aligned} z &= \text{NORM.S.INV}(0.05) \\ &\approx -1.6449 \end{aligned}$$

So a z-score of approximately -1.6449 separates the lower 5% area in the standard normal distribution from the upper 95% area. These z-scores are often called critical z-scores as they are critical boundary values for specific area measures in the standard normal distribution. We now attempt similar text exercises.

? Text Exercise 4.6.2

After sketching the regions described, find z-score(s) that produce the area described in the standard normal distribution.

1. Find the z-score associated with the 65^{th} percentile of the standard normal distribution.

Answer

After producing a sketch indicating the concept of 65^{th} percentile in a standard normal distribution graph (shown below), we notice we are working with a left-tailed region.

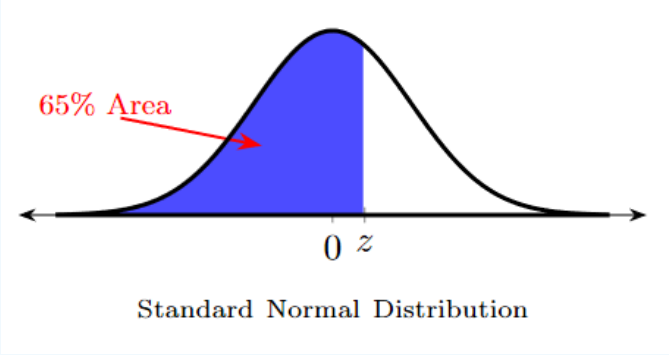


Figure 4.6.10 Standard normal distribution with shaded area

To find the boundary z-score value associated with this left-tailed area measure, we can go directly to our inverse accumulation function with no adjustment.

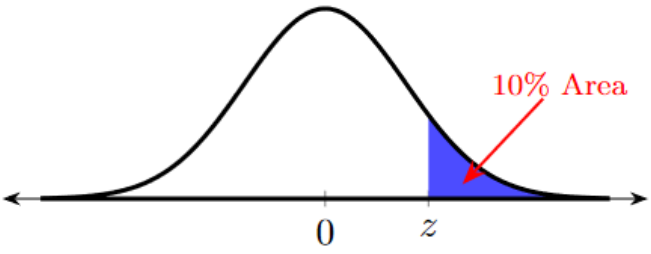
$$\begin{aligned} z &= \text{NORM.S.INV}(0.65) \\ &\approx 0.3853 \end{aligned}$$

Thus, 65% of the standard normal distribution's area is to the left of a z-score of 0.3853. Equivalently, there is a 65% probability of randomly selecting a z-score outcome that is less than 0.3853 in value. We also note that such decimal values for z-scores will occur more frequently in our computation results. We must note the type of value we are computing/measuring and not depend on what the value looks like to control our interpretation. Remember that probabilities are numbers between 0 and 1; whereas, z-scores can be any real number, including the numbers between 0 and 1.

2. Find the z-score so that 10% of the standard normal distribution is above that z value.

Answer

After producing a sketch indicating the described region in the exercise (shown below), we notice we are working with a right-tailed region.



Standard Normal Distribution

Figure 4.6.11: Standard normal distribution with shaded area

We must make complement adjustments in our work using our inverse accumulation function to find the boundary z-score value associated with this right-tailed area measure. We notice that the left-tailed (white) region under our curve must be $1-0.10=0.90=90\%$ of the total area based on our complement rule. We find our boundary z-score value by:

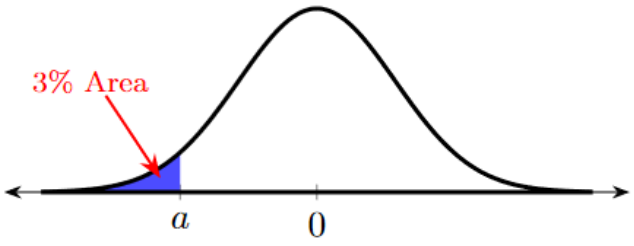
$$\begin{aligned} z &= \text{NORM.S.INV}(0.90) \\ &\approx 1.2816 \end{aligned}$$

Thus, 10% of the standard normal distribution's area is to the right of a z-score of 1.2816.Or equivalently, there is a 10% probability of randomly selecting a z-outcome that is at least 1.2816in value. We should be careful with the syntax here. Notice that entering $1-\text{NORM.S.INV}(0.1)$ does not produce the correct answer.

3. Find the value of a so that $P(z \leq a)=3\%$.

Answer

After producing a sketch indicating the 3% left-tailed region in a standard normal distribution graph (shown below), we go directly to our inverse accumulation function to compute the related z-score labeled with a.



Standard Normal Distribution

Figure 4.6.12 Standard normal distribution with shaded area

$$\begin{aligned} a &= \text{NORM.S.INV}(0.03) \\ &\approx -1.8808 \end{aligned}$$

Thus, 3% of the standard normal distribution's area is to the left of a z-score of -1.8808 Equivalently, there is a 3% probability of randomly selecting a z-score that is most -1.8808in value; stated symbolically, $P(z \leq -1.8808)=3\%$. Note that we would label this as an unusual outcome.

4. Find a for which $P(-a<z<a)=80\%$; that is find the z-scores that capture the central 80% of the standard normal distribution.

Answer

We must adjust our computation work for left-tailed regions After producing a sketch indicating the central 80%region in a standard normal distribution graph (shown below).

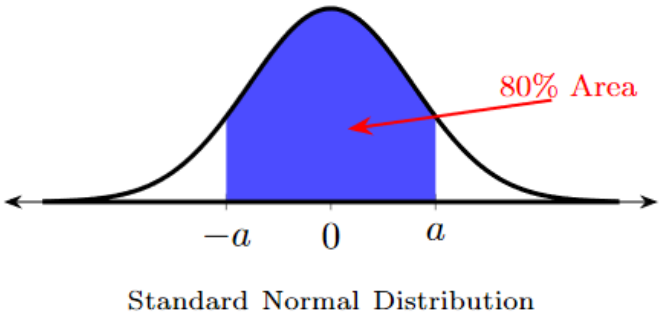


Figure 4.6.13 Standard normal distribution with shaded area

With 80% in the central region, that leaves $100\% - 80\% = 20\%$ area measure for the two tails. This means that there is an area of $\frac{20\%}{2} = 10\%$ in each tail since the standard normal distribution is symmetric. To find the left-boundary z-score value (labeled as $-a$ in the diagram), we use our inverse accumulation function with the 10% left-region area measure.

$$\begin{aligned} z &= \text{NORM.S.INV}(0.10) \\ &\approx -1.2816 \end{aligned}$$

To find the right-boundary z-score value (labeled as a in the diagram), we can use the symmetry of the standard normal curve about the central value 0 to reason that $a = 1.2816$. For extra practice, we can also compute with our inverse accumulation function with a $80\% + 10\% = 90\%$ left-region area measure.

$$\begin{aligned} z &= \text{NORM.S.INV}(0.90) \\ &\approx 1.2816 \end{aligned}$$

Thus, 80% of the standard normal distribution's area is between z-scores of -1.2816 and 1.2816.

We have now found ways to use a technology accumulation function and its inverse to produce various area and scale measures of the standard normal distribution. However, many normal distributions are not the standard normal distribution. We now examine the same ideas for any normal distribution.

Area Measures in Non-standard Normal Distributions

We discuss two methods for finding probabilities as well as the inverse action when working with normal distributions that are not standard normal (the mean is not zero and/or the standard deviation is not one). We will be using both methods on this concept in the remainder of our text, so it is important for us to learn the methods well now before adding more concepts.

Conversion to Standard Normal

As reviewed in Section 4.5, we can convert any normally distributed random variable, x, into the scale of the standard normal variable, z, using our standardization calculation: $z = \frac{x-\mu}{\sigma}$. This implies that we can compute any needed areas and z-values for any normal distribution by using this conversion process first and then applying the concepts from the standard normal distribution.

For example, suppose that the time for various college students to complete a specific task is normally distributed with $\mu = 25$ minutes and $\sigma = 5$ minutes, and we want to know what proportion of the students spent less than 15 minutes to complete the task. We recall that a normal probability distribution is determined by its mean and standard deviation values, allowing us to quickly sketch the distribution, including reasonably accurate scaling of our horizontal axis. Based on the provided information, we graph this non-standard normal distribution in Figure 4.6.14, along with its standardization into the z-distribution.

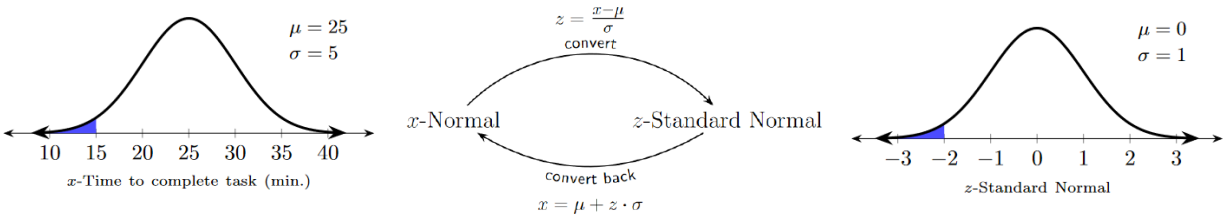


Figure 4.6.14: Standardization of a normal distribution

Often, to save space and sketching time, we do not write the vertical scale on our normal distributions as seen above (we do note that the vertical scaling is different between the two distributions since the horizontal scaling is different, but for our current purposes knowing the difference is not essential). Often, we will place the natural/raw and standardized scaling in the same distribution sketch, as shown in Figure 4.6.15 below.

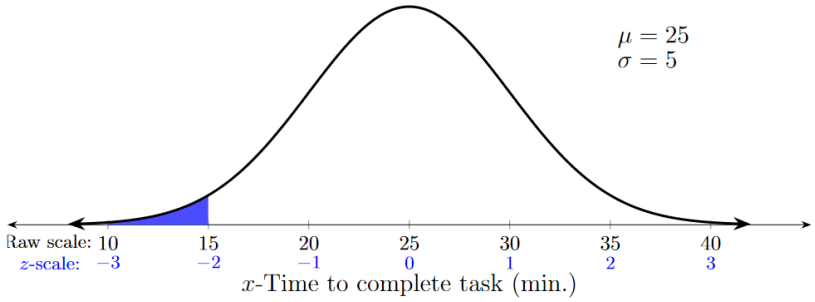


Figure 4.6.15: Standardized scaling on a non-standard normal distribution

Since we are seeking $P(x < 15 \text{ min.})$, we standardize 15 min. to $z = \frac{15-25}{5} = -2$. Intuitively, this means that 15 minutes is 2 standard deviations below the mean. As can be seen in either Figure 4.6.14 or 4.6.15, the area of the region to the left of 15 in the specified normal distribution is the same as the area to the left of -2 in the standard normal distribution; that is, $P(x < 15\text{min}) = P(z < -2)$. Since we can compute $P(z < -2)$ by $= \text{NORM.S.DIST}(-2, 1)$ in our spreadsheet, producing the value 0.0228, we know that $P(x < 15\text{min}) = 0.0228 = 2.28\%$. That is, about 2.28% of those college students spent less than 15 minutes to complete the task.

The key idea here is that we can convert back and forth between any given normal distribution scale and the standard normal distribution scale to handle probability questions related to the non-standard normal distribution. We illustrate one more example below with an "inverse" function problem where we "convert back" to find our needed measures.

Suppose, in the same random variable context, we want to know the time interval the central 80% of those students took to complete the task. As shown in Figure 4.6.16 below, we need to find the scale values, labeled as a and b, that captures the central 80% of the distribution's entire region. Notice that, even though we technically have the same horizontal scale values known to us, we have very little of the x- or z-axes scaled in our sketch as compared to our Figures 4.6.14 and 4.6.15. This is because, initially, we are unsure where these boundary values for the 80% region are exactly located until computed in our later work. Again, a reasonable sense of the figure is essential for answering this "inverse" question.

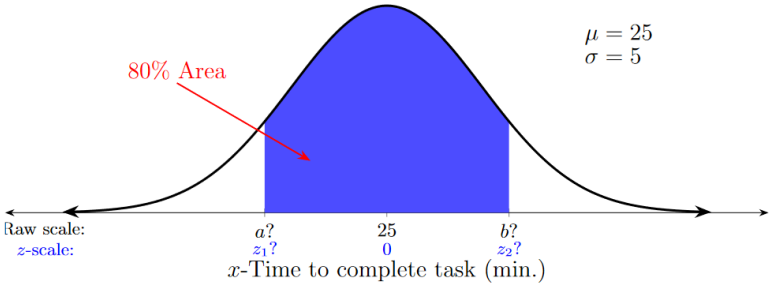


Figure 4.6.16: Inverse conversion to find scale value

We did such "inverse" or "convert back" work above in the standard normal distribution with related left-area measures producing the two results of

$$z_1 = \text{NORM.S.INV}(0.10) \approx -1.2816$$
$$z_2 = \text{NORM.S.INV}(0.90) \approx 1.2816.$$

If we reverse the conversion process, taking these z-scores back to the related x-scores using inverse formula $x = \mu + z \cdot \sigma$, we can produce the related raw scale values of a and b:

$$a = \mu + z_1 \cdot \sigma \approx 25 + (-1.2816) \cdot (5) = 18.592$$
$$b = \mu + z_5 \cdot \sigma \approx 25 + (1.2816) \cdot (5) = 31.408.$$

Finally, thinking about the contextual interpretation of these results, we know that the central 80% of those college students (that is, a large majority of them) took between 18.6 minutes and 31.4 minutes to complete the task. Knowing such information might be useful in planning the time one should allot so that most can complete the task on time.

? Text Exercise 4.6.3

Sketch the distributions described and find the desired value(s).

1. If a random variable x has a normal distribution with $\mu = 18.2$ and $\sigma = 3.4$, find $P(x > 22)$, that is, find the proportion of this distribution that is above 22.

Answer

First, we sketch a diagram of the described normal probability distribution with the standardized scale on the distribution.

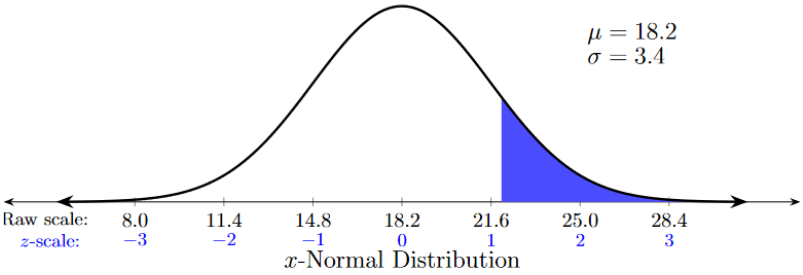


Figure 4.6.17: Normal distribution with shaded area

We note that this region is right-tailed, with a complement left-tailed region. To determine $P(x > 22)$, we must move from $x = 22$ to the related standardized value of $z = \frac{22-18.2}{3.4} \approx 1.1176$. Hence,

$$\begin{aligned} P(x > 22) &= P(z > 1.1176) \\ &= 1 - P(z \leq 1.1176) \\ &= 1 - \text{NORM.S.DIST}(1.1176, 1) \\ &\approx 1 - 0.8681 \\ &= 0.1319 = 13.19\%. \end{aligned}$$

That is, the proportion of this normal distribution with values above 22 is about 13.19%.

2. If a random variable x has a normal distribution with $\mu = 18.2$ and $\sigma = 3.4$, find the 25^{th} percentile value for the distribution.

Answer

Sketching a new diagram of the given normal distribution and given conditions on that distribution, remembering that the 25^{th} percentile is equivalent in meaning to the boundary value that separates the lower 25% of the distribution from the upper 75% :

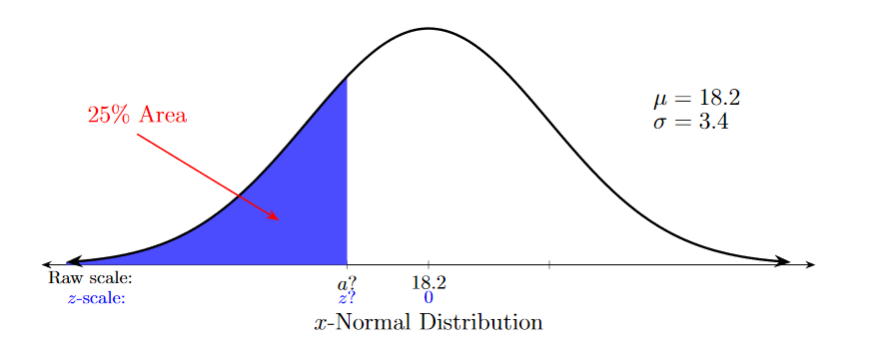


Figure 4.6.18 Normal distribution with shaded area

We seek our shaded region's boundary value a. We must determine the related z-score first through the use of our NORM.S.INV function and then convert that z-score back to our raw-scale score. We first compute our critical z-score by

$$\begin{aligned} z &= \text{NORM.S.INV}(0.25) \\ &\approx -0.6745, \end{aligned}$$

then convert to raw value by

$$\begin{aligned} a &= \mu + z \cdot \sigma \\ &\approx 18.2 + (-0.6745) \cdot 3.4 \\ &= 15.9067. \end{aligned}$$

Hence, the 25^{th} percentile value in this given normal distribution is approximately at the value $x = 15.91$.

3. A soft drink bottler has data that suggest that the amount of drink placed in their 12-ounce cans by a specific bottling machine is normally distributed with $\mu = 12.1$ ounces and $\sigma = 0.5$ ounces (the machine is slightly over-filling on average from designed specifications).
   a. What proportion of cans are under-filled from the labeled amount by more than 1 ounce?
   b. What amount of soft drink in the cans accounts for the central 90% of all cans filled by this specific machine?

Answer

a. Below is our sketch of the situation, noting we are involved with a left-tailed region:

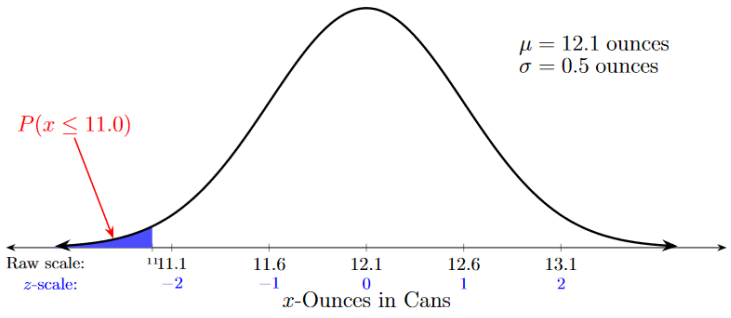


Figure 4.6.19 Normal distribution with shaded area

To find the proportion of under-filled cans by this machine from the labeled amount of 12 ounces by more than 1 ounce, we need to find $P(x < 11)$. First we convert to $z-$ scale:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{12.1 - 11}{0.5} \\ &= -2.20, \end{aligned}$$

then find our area measure in the standard normal distribution:

$$\begin{aligned} P(x < 11) &= P(z < -2.20) \\ &= \text{NORM.S.DIST}(-2.20, 1) \\ &\approx 0.0139 = 1.39\%. \end{aligned}$$

So about 1.39% of the cans are being under-filled by more than one ounce from the desired specifications. That value shows that under-filling by more than one ounce is unusual for this machine.

b. We sketch a diagram of the given information:

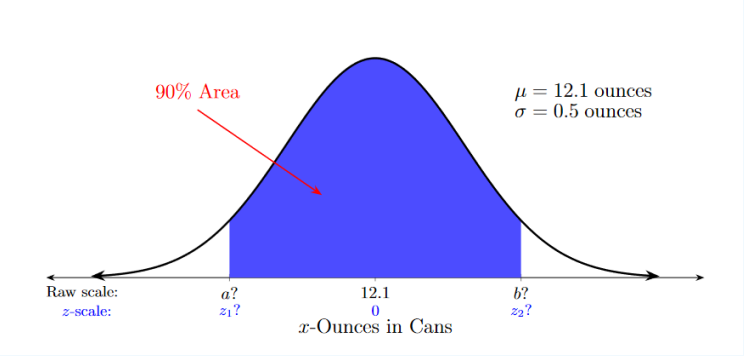


Figure 4.6.20 Normal distribution with shaded area

We seek the boundary values a and b in our raw scaled axis to capture the central 90% of the normal probability distribution. However, we again must first get the related z_1- and z_2-scores through the use of our NORM.S.INV function and then convert them back to our raw-scale score. So, after noticing we have 5% of the area in the white regions of the two tails in our distribution, we compute our two symmetrical critical z-scores by

$$\begin{aligned} \pm z &= \pm\text{NORM.S.INV}(0.05) \\ &\approx \pm 1.6449, \end{aligned}$$

then convert to raw scale by

$$\begin{aligned} a &= \mu + z_1 \cdot \sigma \\ &\approx 12.1 + (-1.6449) \cdot 0.5 \\ &= 11.2776 \end{aligned} \quad \text{and} \quad \begin{aligned} b &= \mu + z_2 \cdot \sigma \\ &\approx 12.1 + (1.6449) \cdot 0.5 \\ &= 12.9224. \end{aligned}$$

Hence, 90% of the cans being filled by this machine have between 11.28 and 12.92 ounces in them.

By working in the standard normal distribution with left-tail regions, we can determine the areas' related z-scale values. These z-scale values can then be "converted back" into the scaled values of the non-standard normal distribution. This back-and-forth conversion work between the x-scale and the z-scale can get tedious, but it is a beneficial strategy for working with normal distributions. We all likely need more practice by doing several homework problems to get reasonable mastery of these ideas. For some of our later work in inferential statistics, this type of conversion work and the meaning of this conversion action will be extremely important.

Nonetheless, in the following subsection, we explore our second method and two new spreadsheet functions that hide/automate this conversion process, allowing us to keep within the natural/raw scale of the given normal probability distribution.

Hidden/Automated Conversion to Standard Normal

We introduce an **accumulation function for any normal distribution**. The name and syntax of this function can vary depending on the technology one uses, but the name of the accumulation function in Excel is NORM.DIST. We note that the spreadsheet function name here only misses the "$.S$" required for the standard normal distribution function. This function requires we provide it with a specific x-scale value in the distribution as well as the mean μ and the standard deviation σ of the normal distribution. In general, the syntax of this accumulation function is $= \mathbf{NORM.DIST(}\boldsymbol{x}\mathbf{\text{-score}, }\boldsymbol{\mu}\mathbf{, }\boldsymbol{\sigma}\mathbf{, TRUE)}$ or the slightly shorter version of $= \mathbf{NORM.DIST(}\boldsymbol{x}\mathbf{\text{-score}, }\boldsymbol{\mu}\mathbf{, }\boldsymbol{\sigma}\mathbf{, 1).}$ The function will return the area of the region to the left of that x-value if we choose the TRUE option. Similar to the standard normal distribution's function, we can enter the digit 1 instead of typing out the word TRUE when using this accumulation function. Also, similarly, if we use FALSE with the function, it returns only the height of the density function at that specific x-value, not an area.

Let us re-examine the example problems from the last section in which we converted back and forth between the normal distribution of interest and the standard normal distribution. Recall our given context in which the time for various college students to complete a specific task is normally distributed with $\mu = 25$ minutes and $\sigma = 5$ minutes. We again ask, what proportion of the students spent less than 15 minutes to complete the task? We graph this in Figure 4.6.21 based on this information. In our second method, we do not include scaling with the related standardized z-scores:

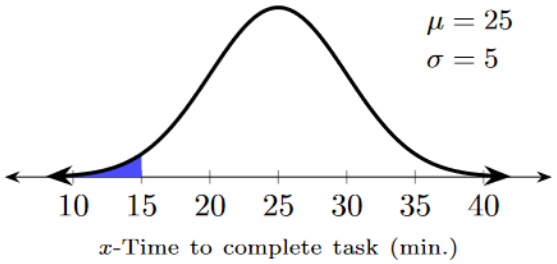


Figure 4.6.21: Standardization of a specific normal distribution

Since we are seeking $P(x < 15 \text{ min.})$, we need to compute the area of the region to the left of 15. In this approach, we use our general normal distribution accumulation function instead of standardizing values. We can compute $P(x < 15)$ through $= \text{NORM.DIST}(15, 25, 5, 1)$ in our spreadsheet producing the value 0.0228. This is the same value we computed using the conversion process. We have found that about 2.28% of those college students spent less than 15 minutes to complete the task.

As with the standard normal distribution function, we must remember that this function always produces only a left-tail area measure. If our regions of interest are central or right-tail regions, adjustments must be made similarly to our previous work.

For one more example, suppose in the context of the college student's time to complete a task, we wish to know the probability of randomly selecting a student who took over 37.5 minutes on the task. We produce a quick sketch again for this question:

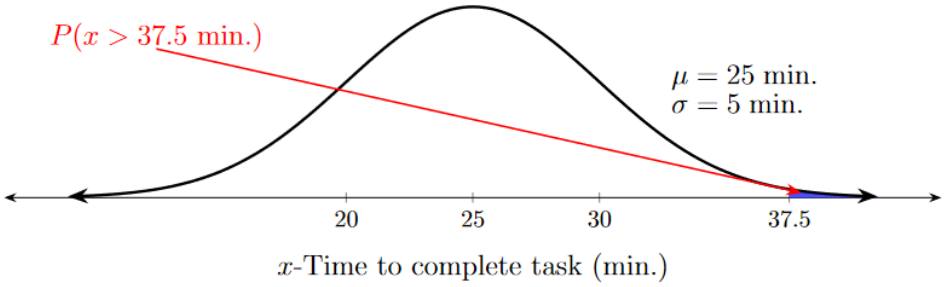


Figure 4.6.22 Probability in a non-standard normal distribution without standardizing

Since we are seeking $P(x > 37.5 \text{ min.})$, our figure shows we need to compute a right-tailed region and use of our complement rule:

$$\begin{aligned} P(x > 37.5) &= 1 - P(x \leq 37.5) \\ &= 1 - \text{NORM.DIST}(37.5, 25, 5, 1) \\ &\approx 1 - 0.9938 = 0.0062 = 0.62\%. \end{aligned}$$

The probability of randomly selecting a student from this group who took over 37.5 minutes on the task is less than 1% and considered an unusual event.

? Text Exercise 4.6.4

Sketch graphs of and determine the designated measures in the following:

1. Find $P(x \geq 122)$ and $P(75 < x < 110)$ in a normal distribution with $\mu = 100$ and $\sigma = 15$.

Answer

First we find $P(x \geq 122)$. Our sketched graph is shown below...we are working in a right-tailed region. We are using an approach that does not require standardization.

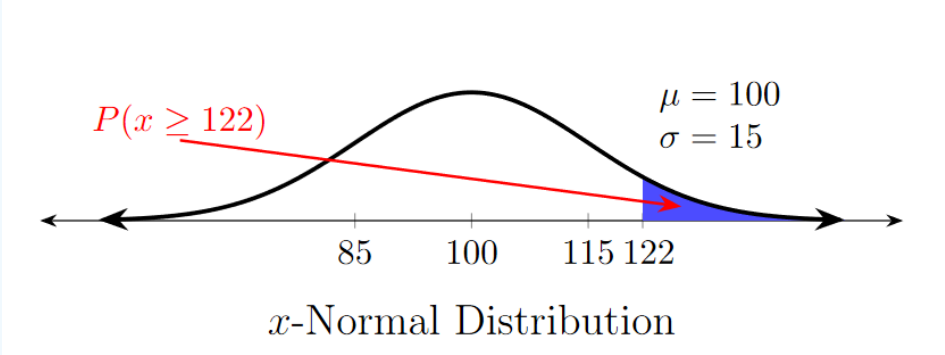


Figure 4.6.23 Normal distribution with shaded region

To find the right-tailed area measure, we make our complement adjustment.

$$\begin{aligned} P(x \geq 122) &= 1 - P(x < 122) \\ &= 1 - \text{NORM.DIST}(122, 100, 15, 1) \\ &\approx 1 - 0.9288 \\ &= 0.0712 = 7.12\% \end{aligned}$$

Remembering our quick check, we notice that the size of the shaded region in the graph seems to align with this proportional measure of 7.12%.Thus, $7.12\%$of the normal distribution's area is to the right of 122.Or equivalently, there is a 7.12%probability of randomly selecting a x-outcome from this distribution that is at least 122in value.

Next, we find $P(75 < x < 110)$. Our sketched graph is shown below; noticing that we are in a central region, we must subtract two left-tailed area measures.

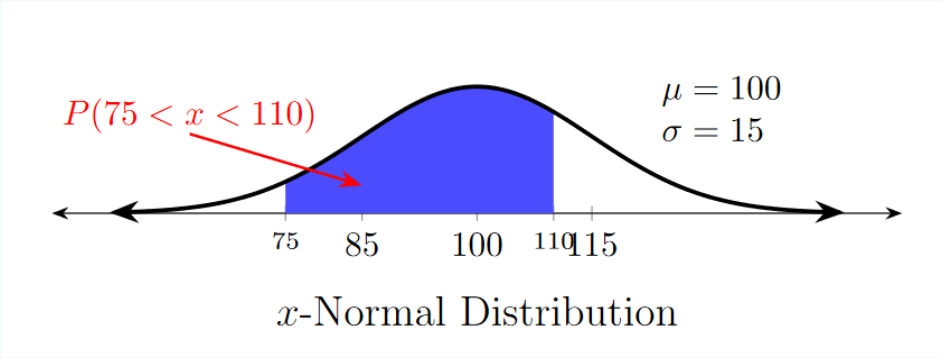


Figure 4.6.24 Normal distribution with shaded region

We subtract two left-tail areas to find the desired region's area measure.

$$\begin{aligned} P(75 < x < 110) &= P(x < 110) - P(x \leq 75) \\ &= \text{NORM.DIST}(110, 100, 15, 1) - \text{NORM.DIST}(75, 100, 15, 1) \\ &\approx 0.7475 - 0.0478 \\ &= 0.6997 = 69.97\% \end{aligned}$$

There is a 69.97%probability of randomly selecting a x-score that is between 75 and 110in value.

2. A soft drink bottler has data that suggest that the amount of drink placed in their 12-ounce cans by a specific bottling machine is normally distributed with $\mu = 12.1$ ounces and $\sigma = 0.5$ ounces. What proportion of cans are under-filled from the labeled amount by more than 1 ounce?

Answer

After carefully reading the context about filling the cans with soft drinks, we produce the volume of soft drink probability distribution graph below.

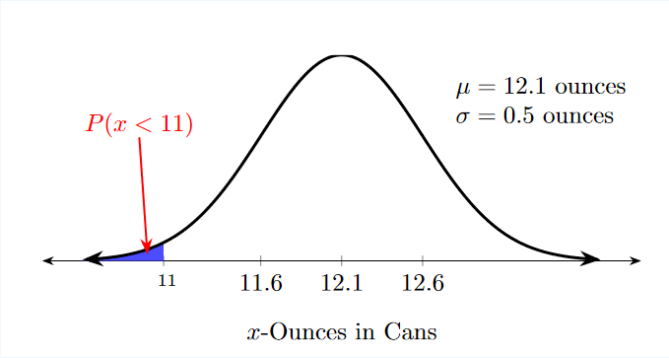


Figure 4.6.25 Normal distribution with shaded region

To determine the proportion of cans that are under-filled from the labeled amount by more than 1 ounce, we find $P(x < 11)$. We compute the area of the left-tail region.

$$\begin{aligned} P(x < 11) &= 1 - \text{NORM.DIST}(11, 12.1, 0.5, 1) \\ &\approx 0.0139 = 1.39\% \end{aligned}$$

Since the probability of randomly selecting a can filled by this machine with less than 11 ounces is less than 5%, such an outcome would be considered unusual. We note that this is the same value we produce using the standardization conversion method.

3. The average consumption of electricity by electric four-door passenger vehicles is believed to be normally distributed with $\mu = 0.346$ kWh per mile and $\sigma = 0.022$ kWh per mile, where kWh stands for kilowatt hour. Is our vehicle considered unusual if we own such an electric vehicle that achieves 0.400 kWh per mile?

Answer

After carefully reading the electricity car context, we produced the electricity consumption probability distribution graph below.

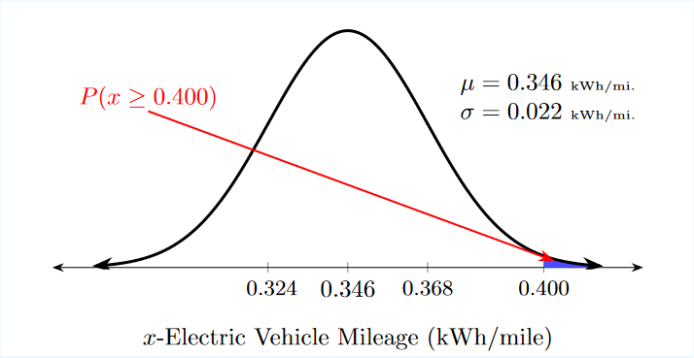


Figure 4.6.26 Normal distribution with shaded region

To determine if our vehicle is getting "unusually" high mileage, we need to determine the probability measure of having a mileage of 0.400 kWh per mile or higher; that is, we need to compute $P(x \geq 0.400$. Per the graphic, even without computation, it appears that the probability is small. However, scales can be deceiving, so we compute the value to have measurement evidence to base our conclusion. As this is a right-tail region, we use our complement adjustment.

$$\begin{aligned} P(x \geq 0.400) &= 1 - P(x < 0.400) \\ &= 1 - \text{NORM.DIST}(0.400, 0.346, 0.022, 1) \\ &\approx 1 - 0.9929 \\ &= 0.0071 = 0.71\% \end{aligned}$$

Since the probability of getting 0.400 kWh per mile or higher is less than 1%, our electric vehicle would be considered unusual. We are getting unusually high mileage compared to similar electric vehicles.

4. Based on data taken from a large group of healthy humans in the United States, human body temperatures seem to be normally distributed with $\mu = 98.3^\circ\text{F}$ and standard deviation$\sigma = 0.92^\circ\text{F}$. If a local hospital uses 100.5°F as the lowest temperature indicating a likely fever and illness, what percentage of healthy humans will be classified as ill by this hospital?

Answer

We produce the following graph of the normal distribution of healthy body temperatures. Noting a person would be considered feverish by this hospital if they have body temperatures over 100.5°F, shown in the shaded region of the probability distribution. To compute this region, we must use the complement.

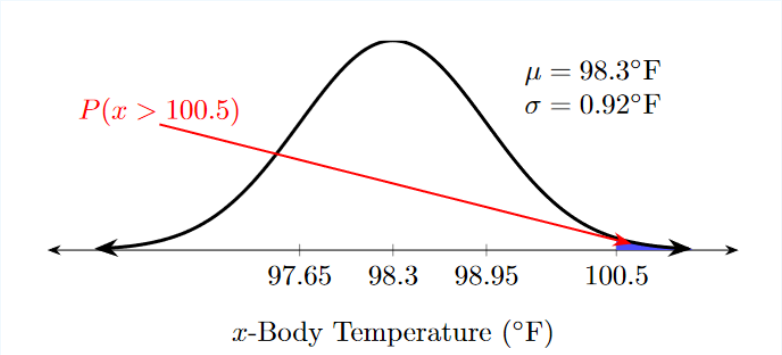


Figure 4.6.27: Normal distribution with shaded region

$$\begin{aligned} P(x \geq 100.5) &= 1 - P(x < 100.5) \\ &= 1 - \text{NORM.DIST}(100.5, 98.3, 0.92, 1) \\ &\approx 1 - 0.9916 \\ &= 0.0084 = 0.84\% \end{aligned}$$

Less than 1% of healthy individuals will be classified by this hospital as feverish when they are not ill. The hospital will not likely run into this situation very often.

There will also be occasions in which we need to reverse the area/probability process above. Given the description of a region and its area measure, can we find the horizontal scale measure(s) that serve as boundary(ies) of the described region? We can use our standardization process, but there are functions for this inverse process that take care of the calculations for us and leave us within the raw/natural scale of the situation.

We return to one of our earlier questions: what are the x-scores in the normal distribution of student times for completing the specific task that produces the central 80% region of that distribution? This time, as we did in an earlier analysis, we want to avoid converting to standard normal distribution measures on our axis. As shown in Figure 4.6.28, we need to find the scale values labeled as a and b in this normal distribution that captures the central 80% of the distribution's region.

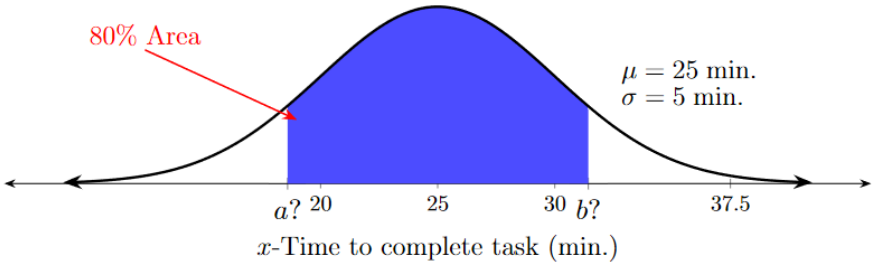


Figure 4.6.28: Inverse conversion to find scale value(s) in a non-standard normal distribution

We are saved from manually doing all the conversion work by a new inverse function, NORM.INV, that behaves similarly to our already familiar NORM.S.INV function from the standard normal distribution. If given any left-tail region's area measure, this function will compute the associated right boundary x-scale value forming that region, provided the mean μ and standard deviation σ are both known. This function has the syntax $=\text{NORM.INV}(\text{left-tail area measure between } \mathbf{0} \text{ and } \mathbf{1}, \boldsymbol{\mu}, \boldsymbol{\sigma})$. Again, we emphasize that the function provides values only for left-tailed regions.

For the boundary value a in our diagram, we note a left-tail area measure of $10\% = 0.10$. We compute in our spreadsheet:

$$\begin{aligned} a &= \text{NORM.INV}(\text{left area measure}, \mu, \sigma) \\ &= \text{NORM.INV}(0.10, 25, 5) \\ &\approx 18.5922 \text{ min.} \end{aligned}$$

An x-score of approximately 18.5922 separates the lower 10% area in the given normal distribution from the upper 90% area. We also need to determine our right boundary value b in Figure 4.6.28, which has a left-tail area measure of 90%.

$$\begin{aligned} b &= \text{NORM.INV}(\text{left area measure}, \mu, \sigma) \\ &= \text{NORM.INV}(0.90, 25, 5) \\ &\approx 31.4078 \text{ min.} \end{aligned}$$

As a completed result, the central 80% of the students had completion times for the task between approximately 18.6 and 31.4 minutes.

We can compare these results with our earlier work (which included inverse conversion work from the standard normal distribution) to see that they are the same. We now attempt similar text exercises involving inverse distribution methods.

Text Exercise 4.6.5

After sketching the regions described, find x-score(s) that produce the area measures described in the normal distribution.

1. When designing a building, a common requirement is to design for 95% of the population that will be using that building. To be safe as well as cost effective, and since men on average are taller than women, a building's doorways are to be designed so that all but the tallest 5% of men can walk through the doorway without having to stoop. If the heights of men are normally distributed with a mean of 161.29cm with standard deviation of 8.3 cm, determine the design height needed for doorways.

Answer

Our sketch of the height distribution for men is shown below, with shading indicating the separation between the lower 95% and upper 5% in the distribution.

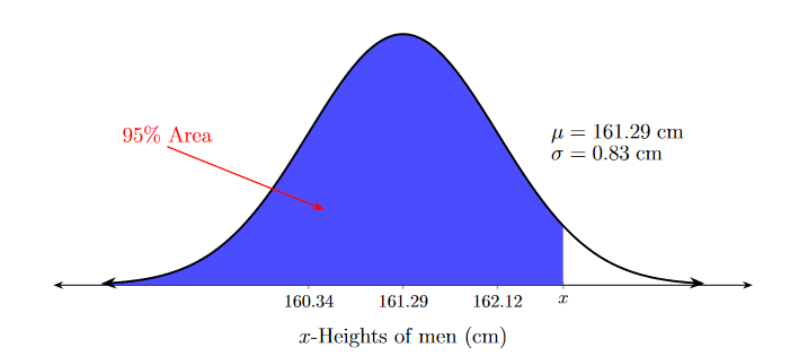


Figure 4.6.29 Normal distribution with shaded region

To find the boundary x-height value associated with the left-tailed 0.95 area value, we can go directly to our inverse accumulation function.

$$\begin{aligned} x &= \text{NORM.INV}(0.95, 161.29, 0.83) \\ &\approx 162.6552 \end{aligned}$$

Thus, 95% of men have heights below 162.66 cm and 5% have heights above. The building should be designed with doorways that have heights of at least 162.66cm to fit the design specifications.

2. The average consumption of electricity by electric four-door passenger vehicles is believed to be normally distributed with $\mu = 0.346$ kWh per mile with $\sigma = 0.022$ kWh per mile, where kWh stands for kilowatt hour. What is the central 90% expected average consumption for these types of electric vehicles?

Answer

After sketching our normal distribution (shown below), we are seeking the two boundary values of a and b that separate the central 90% of our distribution.

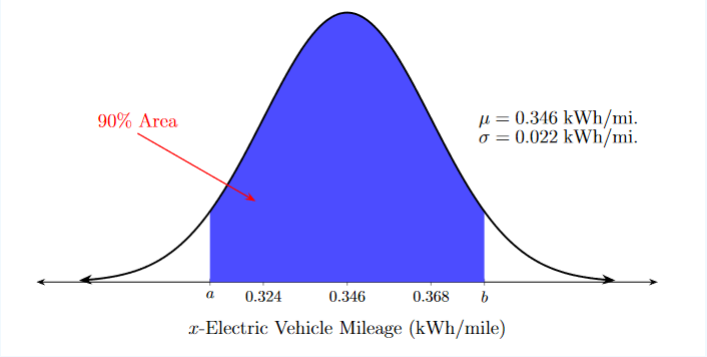


Figure 4.6.30 Normal distribution with shaded region

To find the left boundary value a, we use our inverse accumulation function with the 5% left-region area measure.

$$\begin{aligned} a &= \text{NORM.INV}(0.05, 0.346, 0.022) \\ &\approx 0.3098 \end{aligned}$$

To find the right-boundary b, we again use the inverse accumulation function with a $90\% + 5\% = 95\%$ total left-region area measure.

$$\begin{aligned} b &= \text{NORM.INV}(0.95, 0.346, 0.022) \\ &\approx 0.3822 \end{aligned}$$

Thus, the central 90% average consumption of electricity by these electric vehicles is expected to be between 0.310 and 0.382kWh per mile.

3. A tire company is about to begin large-scale manufacturing of a new tire made of newly developed materials. The tire's tread life has been tested, the research team found the tread life in miles produced a normal distribution with $\mu = 72,000$ miles and $\sigma = 7,000$ miles. The company must develop a consumer warranty policy and only wants to replace tires that do not last sufficiently to tested expectations. The company decides to only set the mileage warranty to cover the lowest 5% of their tires. What is the mileage number they will need to place on the warranty policy?

Answer

We sketch the described tire-life distribution (shown below). Next, we go directly to our inverse accumulation function to compute the related x-mileage value for the boundary value establishing the lowest 5% of the distribution's area.

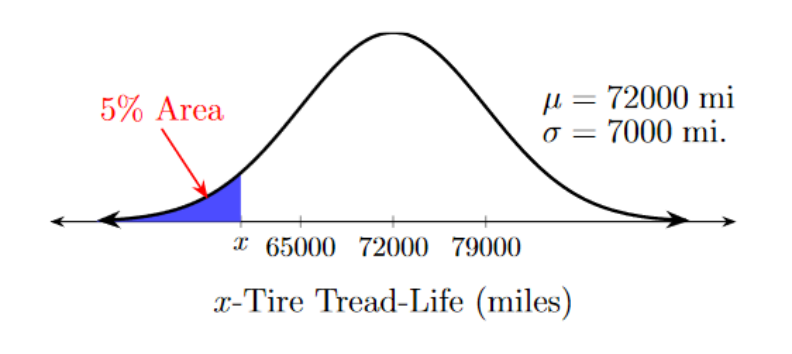


Figure 4.6.31: Normal distribution with shaded region

$$\begin{aligned} x &= \text{NORM.INV}(0.05, 72000, 7000) \\ &\approx 60,486.02 \end{aligned}$$

Thus, 5% of the tires can be expected to last less than $60,486$ miles. The tire company should set their replacement warranty value near this value, likely at $60,000$miles just to round to an easier value for customers.

4. Established in 1946, Mensa, currently a global community of around $150,000$people, requires individuals first score in the upper 2% (in relation to the general population) on an IQ test before being considered for membership. If the general population produces normally distributed scores with $\mu = 100$ points and $\sigma = 15$ points on the IQ test, what must we score on the exam to be considered for membership in Mensa?

Answer

Our sketch of the distribution for IQ scores is shown below, with shading indicating the upper 2% region of the scores.

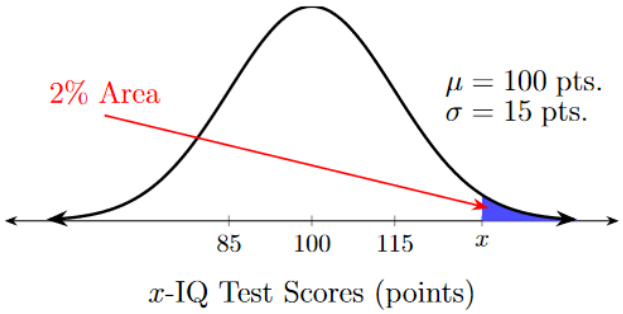


Figure 4.6.32 Normal distribution with shaded region

To find the boundary x-height value associated with the right-tailed 0.02 area value, we can apply complement action ($1 - 0.02 = 0.98$) within the use of the inverse accumulation function.

$$\begin{aligned} x &= \text{NORM.INV}(0.98, 100, 15) \\ &\approx 130.8062 \end{aligned}$$

We must score at least 130.8points on the IQ test to be within the top 2%.

We have now found ways, using another technology accumulation function and its inverse, to be able to produce various area or x-scale measures of any x-normal distribution. These new methods eliminated in our work the converting back and forth between a general normal distribution and the standard normal distribution. Yes, the above technology does make our work less intense by hiding the conversion work (in the programing of the functions the conversion work is actually still happening). This is a blessed simplification for us humans as we prefer to eliminate calculation work when possible. However, we again warn that applying the conversion process is necessary in some of our future work, so we should practice both approaches.

Summary

We now have the tools to answer practically any probability related question tied to normal distributions. Our technology's accumulation functions will produce accurate measures of left-region areas of all types of normal distributions. The inverse functions will allow us to find scale measures tied to given regions of a normal probability distribution. In the future, we will also examine similar accumulation functions for other common probability distributions, such as the t- and χ^2-distributions.

CHAPTER OVERVIEW

5: Sampling Distributions

5.1: Introduction to Sampling Distributions

Learning Objectives

- Define and construct probability distributions of sample statistics with simple random sampling
- Define and construct sampling distributions of sample statistics
- Define and give examples of unbiased estimators
- Explore the impact sample size has on the sampling distributions of a given population
- Represent sampling distributions as continuous random variables

Review and Preview

In the previous chapter, we began our study of random variables by briefly connecting them with inferential and sample statistics. This chapter studies sample statistics as random variables, paying close attention to probability distributions. Recall for each random variable, an underlying random experiment will be conducted. A value is assigned, measured, or computed for each possible outcome of this random experiment. These are the values the random variable is said to take on, and the probability of the values occurring is determined, forming a probability distribution. If the random variable is discrete, each value has a specific probability. If the random variable is continuous, a range of values has a probability based on the area under the probability density function. We can compute the expected value (mean), variance, and standard deviation to help describe the random variable.

Suppose we wanted to know the average height of an American. It is impossible to measure the height of everyone in the country, so we decided to measure the heights of 10,000 randomly selected Americans and found that the average height of these 10,000 people is 68.4 inches. We cannot conclude that the average height of all Americans is 68.4 inches. After all, we may have selected the 10,000 tallest people in the country, or perhaps 8,000 were well below the average height. Even if our sampling method is unbiased, there is still the possibility that we obtained a sample mean far away from the population mean by pure chance. Even if the sample mean is close to the population mean, it likely is not the same. Do we need to measure more heights? Can 10,000 people accurately represent the whole country? To understand the average height of an American, we need to determine the probability that our sample mean is not accurate. What is the probability that our sample mean is off by more than 1 inch? What is the probability that it is off by less than 0.1 inches? To answer these questions, we need to think of sampling as a random variable.

Now that we have looked at the basics of random variables and have an example in mind, we study sample statistics as random variables in depth. We are interested in learning the characteristics of a population (parameters). Studying the entire population may be impossible, too expensive, or time-consuming, so we study a sample and compute a statistic to estimate the parameter. Ideally, the sample is representative of the population; it does not misrepresent the population, and the statistic is close to the parameter. We cannot guarantee such a sample, but by choosing our sample randomly, we can ensure that any bias in the sample is due to random chance. For our initial purposes, we work within the context of simple random sampling. We must decide how large of a sample to use; we denote the sample size as n. The random experiment is defined as randomly selecting n objects from the population. The sample space of our random experiment consists of all the possible samples of size n taken from the population of size N. There are ${}_NC_n$ possible samples. Each sample is assigned a value by computing the sample statistic of interest. These possible values, along with their probabilities, form the probability distribution of the sample statistic under simple random sampling. The questions of interest are: what values can the sample statistic take on, and what are the probabilities?

Constructing Probability Distributions of Sample Statistics: Proportions

We cannot know the values and their associated probabilities without studying the entire population. However, we can build solid intuition by studying small populations, where we can exhaustively study the probability distributions. We will be able to generalize in the following sections.

Consider the family of five that has been with us throughout the text. Adam and Betsy have three children: Cathy, Damon, and Erin. This family will serve as our initial population of interest. We will study several different characteristics of this family. The first step is to decide how many family members we want to sample. Indeed, it is unnecessary to sample in this situation, but this process builds our intuition and understanding of the topic. Sampling 3 family members will be perfect for our example.

Text Exercise 5.1.1

1. Determine the number of samples and then list all possible samples of size $n = 3$ from the population of Adam (A), Betsy (B), Cathy (C), Damon (D), and Erin (E).

Answer

We are selecting 3 family members from a family of 5. The order in which we select them does not matter; therefore, there are ${}_5C_3 = 10$ possible samples.

ABC	ABD	ABE	ACD	ACE
ADE	BCD	BCE	BDE	CDE

Notice the pattern of our labelling. Adopting such a method ensures all possible samples are listed easily.

2. Recall that Adam, Betsy, and Cathy all wear glasses. The population proportion p of family members that wear glasses is $p = \frac{3}{5} = 60\%$. For each possible sample of size 3, determine the sample proportion $\hat{p}$.

Answer

We must compute the sample proportion for each of the 10 samples from part one of this text exercise.

Table 5.1.1: All possible samples and their sample proportions

Sample	$\hat{p}$	Sample	$\hat{p}$
ABC	$\frac{3}{3} = 1$	ADE	$\frac{1}{3}$
ABD	$\frac{2}{3}$	BCD	$\frac{2}{3}$
ABE	$\frac{2}{3}$	BCE	$\frac{2}{3}$
ACD	$\frac{2}{3}$	BDE	$\frac{1}{3}$
ACE	$\frac{2}{3}$	CDE	$\frac{1}{3}$

Note that none of the possible sample proportion values equal the population proportion. Recall from our first text exercise with this family that this is true for all possible sample sizes in this particular context. Still, it is not necessarily true for others (see the last part of the referenced exercise for a refresher).

3. We are considering all the samples of size 3 and their sample proportions and have not conducted a random sampling to produce one of these possibilities. We are developing our understanding of the sample proportion $\hat{p}$ as a random variable. There are three possible values that $\hat{p}$ takes on: $\frac{1}{3}$, $\frac{2}{3}$, and 1. Our task now is to fully understand $\hat{p}$ as a random variable and construct its probability distribution.

Answer

Since we are conducting a simple random sampling of size 3 from the family of 5, each sample is equally probable. We can determine the probabilities of our random variable $\hat{p}$ using the classical approach to probability.

Table 5.1.2 Probability distribution of sample proportions

$\hat{p}$	$P(\hat{p})$
$\frac{1}{3}$	$\frac{3}{10}$
$\frac{2}{3}$	$\frac{6}{10} = \frac{3}{5}$
1	$\frac{1}{10}$

4. Determine the expected value, variance, and standard deviation of our random variable $\hat{p}$.

Answer

Table 5.1.3 Computation table

$\hat{p}$	$P(\hat{p})$	$\hat{p} \cdot P(\hat{p})$	$\left(\hat{p}-\mu_{\hat{p}}\right)^2 \cdot P(\hat{p})$
$\frac{1}{3}$	$\frac{3}{10}$	$\frac{1}{3} \cdot \frac{3}{10}=\frac{1}{10}$	$\left(\frac{1}{3}-\frac{3}{5}\right)^2 \cdot \frac{3}{10}=\frac{48}{2250}$
$\frac{2}{3}$	$\frac{6}{10}=\frac{3}{5}$	$\frac{2}{3} \cdot \frac{6}{10}=\frac{4}{10}=\frac{2}{5}$	$\left(\frac{2}{3}-\frac{3}{5}\right))^2 \cdot \frac{3}{5}=\frac{3}{1125}$
1	$\frac{1}{10}$	$1 \cdot \frac{1}{10}=\frac{1}{10}$	$\left(1-\frac{3}{5}\right))^2 \cdot \frac{1}{10}=\frac{2}{125}$
$\mu_{\hat{p}}=\mathrm{E}(\hat{p})=\frac{1}{10}+\frac{4}{10}+\frac{1}{10}=\frac{6}{10}=\frac{3}{5}$			
$\sigma_{\hat{p}}^2=\mathrm{Var}(\hat{p})=\frac{48}{2250}+\frac{6}{2250}+\frac{36}{2250}=\frac{90}{2250}=\frac{1}{25}$			
$\sigma_{\hat{p}}=\sqrt{\mathrm{Var}(\hat{p})}=\sqrt{\frac{1}{25}}=\frac{1}{5}$			

Note that $\mu_{\hat{p}}=\frac{3}{5}=p$. even though none of the possible sample proportions are equal to the population proportion, the expected value of the probability distribution of sample proportions is the population proportion.

When considering probability distributions of sample statistics (for example, sample proportions in our previous exercise), we use subscripts to indicate the sample statistic. This visual reminder provides a simple method of improving clarity and reducing the risk of errors. Reading symbolic expressions with meaning is an important skill to maintain and develop, especially at this point in the course.

Constructing Probability Distributions of Sample Statistics: Means

To continue our exploration, we consider additional data regarding our family of five (presented below): the number of states visited by each person. When considering the characteristic of needing eyeglasses, a qualitative variable, each member of our population either possessed the characteristic or did not. Proportions and modes would be natural sample statistics to consider. On the other hand, the number of states visited is a quantitative variable, so we have many more options to consider for sample statistics. We shall consider some in the following text exercises.

Table 5.1.4: Number of States Visited

Family Member	Number of States Visited
Adam	20
Betsy	30
Cathy	15
Damon	12
Erin	5

? Text Exercise 5.1.2

1. Construct the probability distribution of sample means using a sample size of 3.

Answer

For each of the 10 samples, we must compute the sample mean. We do so by filling out a table.

Table 5.1.5 All possible samples and their sample means

Sample	Sample Data	$\bar{x}$	Sample	Sample Data	$\bar{x}$
ABC	20, 30, 15	$\frac{20+30+15}{3}=\frac{65}{3}$	ADE	20, 12, 5	$\frac{20+12+5}{3}=\frac{37}{3}$

Sample	Sample Data	$\bar{x}$	Sample	Sample Data	$\bar{x}$
ABD	20, 30, 12	$\frac{20+30+12}{3}=\frac{62}{3}$	BCD	30, 15, 12	$\frac{30+15+12}{3}=\frac{57}{3}=1$
ABE	20, 30, 5	$\frac{20+30+5}{3}=\frac{55}{3}$	BCE	30, 15, 5	$\frac{30+15+5}{3}=\frac{50}{3}$
ACD	20, 15, 12	$\frac{20+15+12}{3}=\frac{47}{3}$	BDE	30, 12, 5	$\frac{30+12+5}{3}=\frac{47}{3}$
ACE	20, 15, 5	$\frac{20+15+5}{3}=\frac{40}{3}$	CDE	15, 12, 5	$\frac{15+12+5}{3}=\frac{32}{3}$

Table 5.1.6 Probability distribution of sample means

$\bar{x}$	$P(\bar{x})$
$\frac{32}{3}$	$\frac{1}{10}$
$\frac{37}{3}$	$\frac{1}{10}$
$\frac{40}{3}$	$\frac{1}{10}$
$\frac{47}{3}$	$\frac{2}{10}=\frac{1}{5}$
$\frac{50}{3}$	$\frac{1}{10}$
$\frac{55}{3}$	$\frac{1}{10}$
$\frac{57}{3}$	$\frac{1}{10}$
$\frac{62}{3}$	$\frac{1}{10}$
$\frac{65}{3}$	$\frac{1}{10}$

2. Determine the expected value, variance, and standard deviation of our random variable $\bar{x}$. Compare the expected value of the probability distribution $\mu_{\bar{x}}$ with population mean μ.

Answer

Table 5.1.7 Table of computations

$\bar{x}$	$P(\bar{x})$	$\bar{x}\cdot P(\bar{x})$	$(\bar{x}-\mu_{\bar{x}})^2\cdot P(\bar{x})$
$\frac{32}{3}$	$\frac{1}{10}$	$\frac{32}{3}\cdot\frac{1}{10}=\frac{32}{30}$	$\left(\frac{32}{3}-\frac{82}{5}\right)^2\cdot\frac{1}{10}=\frac{7396}{2250}$
$\frac{37}{3}$	$\frac{1}{10}$	$\frac{37}{3}\cdot\frac{1}{10}=\frac{37}{30}$	$\left(\frac{37}{3}-\frac{82}{5}\right)^2\cdot\frac{1}{10}=\frac{3721}{2250}$
$\frac{40}{3}$	$\frac{1}{10}$	$\frac{40}{3}\cdot\frac{1}{10}=\frac{40}{30}$	$\left(\frac{40}{3}-\frac{82}{5}\right)^2\cdot\frac{1}{10}=\frac{2116}{2250}$
$\frac{47}{3}$	$\frac{2}{10}=\frac{1}{5}$	$\frac{47}{3}\cdot\frac{2}{10}=\frac{94}{30}$	$\left(\frac{47}{3}-\frac{82}{5}\right)^2\cdot\frac{2}{10}=\frac{242}{2250}$
$\frac{50}{3}$	$\frac{1}{10}$	$\frac{50}{3}\cdot\frac{1}{10}=\frac{50}{30}$	$\left(\frac{50}{3}-\frac{82}{5}\right)^2\cdot\frac{1}{10}=\frac{16}{2250}$

$\bar{x}$	$P(\bar{x})$	$\bar{x} \cdot P(\bar{x})$	$(\bar{x} - \mu_{\bar{x}})^2 \cdot P(\bar{x})$
$\frac{55}{3}$	$\frac{1}{10}$	$\frac{55}{3} \cdot \frac{1}{10} = \frac{55}{30}$	$\left(\frac{55}{3} - \frac{82}{5}\right)^2 \cdot \frac{1}{10} = \frac{841}{2250}$
$\frac{57}{3}$	$\frac{1}{10}$	$\frac{57}{3} \cdot \frac{1}{10} = \frac{57}{30}$	$\left(\frac{57}{3} - \frac{82}{5}\right)^2 \cdot \frac{1}{10} = \frac{1521}{2250}$
$\frac{62}{3}$	$\frac{1}{10}$	$\frac{62}{3} \cdot \frac{1}{10} = \frac{62}{30}$	$\left(\frac{62}{3} - \frac{82}{5}\right)^2 \cdot \frac{1}{10} = \frac{4096}{2250}$
$\frac{65}{3}$	$\frac{1}{10}$	$\frac{65}{3} \cdot \frac{1}{10} = \frac{65}{30}$	$\left(\frac{65}{3} - \frac{82}{5}\right)^2 \cdot \frac{1}{10} = \frac{6241}{2250}$
$\mu_{\bar{x}} = \mathrm{E}(\bar{x}) = \frac{32}{30} + \frac{37}{30} + \ldots + \frac{65}{30} = \frac{492}{30} = \frac{82}{5} = 16.4$			
$\sigma_{\bar{x}}^2 = \mathrm{Var}(\bar{x}) = \frac{7396}{2250} + \frac{3721}{2250} + \ldots + \frac{6241}{2250} = \frac{26,190}{2250} = \frac{291}{25} = 11.64$			
$\sigma_{\bar{x}} = \sqrt{\mathrm{Var}(\bar{x})} = \sqrt{\frac{291}{25}} \approx 3.4117$			

To compute the population mean μ, we have $\frac{20+30+15+12+5}{5} = \frac{82}{5} = 16.4$ meaning $\mu_{\bar{x}} = \mu$. The average of all the sample means is the same as the population mean.

Constructing Probability Distributions of Sample Statistics: Range

? Text Exercise 5.1.3

1. Construct the probability distribution of sample ranges using a sample size of 3.

Answer

For each of the 10 samples, we must compute the sample range range. We again do so by filling out a table.

Table 5.1.8 All possible samples and their sample ranges

Sample	Sample Data	Sample Range (range)	Sample	Sample Data	Sample Range (range)
ABC	20, 30, 15	15	ADE	20, 12, 5	15
ABD	20, 30, 12	18	BCD	30, 15, 12	18
ABE	20, 30, 5	25	BCE	30, 15, 5	25
ACD	20, 15, 12	8	BDE	30, 12, 5	25
\(ACE)	20, 15, 5	15	CDE	15, 12, 5	10

Table 5.1.9 Probability distribution of sample ranges

range	$P(\text{range})$
8	$\frac{1}{10}$
10	$\frac{1}{10}$
15	$\frac{3}{10}$
18	$\frac{2}{10} = \frac{1}{5}$

range	$P(\text{range})$
25	$\frac{3}{10}$

2. Determine our random variable's expected value, variance, and standard deviation range. Compare the expected value of the probability distribution $\mu_{\bar{x}}$ with the population range.

Answer

Table 5.1.10 Table of computations

range	$P(\text{range})$	range $\cdot P(\text{range})$	$(\text{range}-\mu_{\text{range}})^2 \cdot P(\text{range})$
8	$\frac{1}{10}$	$8 \cdot \frac{1}{10} = \frac{8}{10} = \frac{4}{5}$	$\left(8-\frac{87}{5}\right)^2 \cdot \frac{1}{10} = \frac{2209}{250}$
10	$\frac{1}{10}$	$10 \cdot \frac{1}{10} = \frac{10}{10} = 1$	$\left(10-\frac{87}{5}\right)^2 \cdot \frac{1}{10} = \frac{1369}{250}$
15	$\frac{3}{10}$	$15 \cdot \frac{3}{10} = \frac{45}{10} = \frac{9}{2}$	$\left(15-\frac{87}{5}\right)^2 \cdot \frac{3}{10} = \frac{432}{250}$
18	$\frac{2}{10} = \frac{1}{5}$	$18 \cdot \frac{2}{10} = \frac{36}{10} = \frac{18}{5}$	$\left(18-\frac{87}{5}\right)^2 \cdot \frac{2}{10} = \frac{18}{250}$
25	$\frac{3}{10}$	$25 \cdot \frac{3}{10} = \frac{75}{10} = \frac{15}{2}$	$\left(25-\frac{87}{5}\right)^2 \cdot \frac{3}{10} = \frac{4332}{250}$
$\mu_{\text{range}} = \text{E(range)} = \frac{4}{5}+1+\ldots+\frac{15}{2} = \frac{174}{10} = \frac{87}{5} = 17.4$			
$\sigma^2_{\text{range}} = \text{Var(range)} = \frac{2209}{250}+\frac{1369}{250}+\ldots+\frac{4332}{250} = \frac{8360}{250} = \frac{836}{25} = 33.44$			
$\sigma_{\text{range}} = \sqrt{\text{Var(range)}} = \sqrt{\frac{207}{10}} \approx 5.7827$			

To compute the population range, we have $30-5 = 25$ meaning $\mu_{\text{range}} \neq \text{Population Range}$. The average of the sample ranges is not the same as the population range. We might have expected this. A sample range could equal the population range if it includes the largest and smallest values. However, for most samples, the sample range will be smaller than the population range because the largest and smallest data values will not be included in the sample. Moreover, the sample range can never exceed the population range. Therefore, the average of the sample ranges is an average of numbers that are never larger than the population range, making the average smaller. This reasoning implies that for any sufficiently varied population, the average of the sample ranges will be less than the population range. Contrast this with the concept of a sample mean.

Probability Distributions of Sample Statistics and Sample Size

In each of our previous examples, we used the sample size $n = 3$. We could have sampled a single person $(n = 1)$ or any number less than 5. If n were to be 5 in our context, $n = N$, we would be studying the entire population, not sampling from it, and then computing the parameter rather than estimating it. We now explore what happens to the probability distributions as we change the sample size while remaining in the same population. Again, we will consider the proportion of family members who wear glasses. The probability distribution of sample proportions for $n = 1, 2, 3$, and 4 and the expected value, variance, and standard deviation are reported in the table below.

Table 5.1.11: Probability Distribution of Sample Proportions for $n = 1, 2, 3$, and 4

	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$\mu_{\hat{p}}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$
$\sigma^2_{\hat{p}}$	0.24	0.09	0.04	0.015

	$n=1$	$n=2$	$n=3$	$n=4$
$\sigma_{\hat{p}}$	≈ 0.4899	0.3	0.2	≈ 0.1225

Students struggling to understand these probability distributions and their construction are encouraged to verify the results above by first building each probability distribution and then computing each measure . Click to check your work.

Probability distribution of sample proportions $n=1$

Table 5.1.12 Probability Distribution of Sample Proportions $n=1$ (5 Samples)

$n=1$ **(5 Samples)**	
$\hat{p}$	$P(\hat{p})$
0	$\frac{2}{5}$
1	$\frac{3}{5}$
$\mu_{\hat{p}}=\frac{3}{5}$	
$\sigma^2_{\hat{p}}=0.24$	
$\sigma_{\hat{p}}\approx 0.4899$	

Probability distribution of sample proportions $n=2$

Table 5.1.13 Probability Distribution of Sample Proportions $n=2$ (10 Samples)

$n=2$ **(10 Samples)**	
$\hat{p}$	$P(\hat{p})$
0	$\frac{1}{10}$
$\frac{1}{2}$	$\frac{6}{10}$
1	$\frac{3}{10}$
$\mu_{\hat{p}}=\frac{3}{5}$	
$\sigma^2_{\hat{p}}=0.09$	
$\sigma_{\hat{p}}=0.3$	

Probability distribution of sample proportions $n=3$

Table 5.1.14 Probability Distribution of Sample Proportions $n=3$ (10 Samples)

$n=3$ **(10 Samples)**	
$\hat{p}$	$P(\hat{p})$
$\frac{1}{3}$	$\frac{3}{10}$
$\frac{2}{3}$	$\frac{6}{10}=\frac{3}{5}$
1	$\frac{1}{10}$

$n = 3$ **(10 Samples)**	
$\hat{p}$	$P(\hat{p})$
$\mu_{\hat{p}} = \frac{3}{5}$	
$\sigma^2_{\hat{p}} = 0.04$	
$\sigma_{\hat{p}} = 0.2$	

Probability distribution of sample proportions $n = 4$

Table 5.1.15 Probability Distribution of Sample Proportions $n = 4$ (5 Samples)

$n = 4$ **(5 Samples)**	
$\hat{p}$	$P(\hat{p})$
$\frac{1}{2}$	$\frac{6}{10}$
$\frac{3}{4}$	$\frac{4}{10}$
$\mu_{\hat{p}} = \frac{3}{5}$	
$\sigma^2_{\hat{p}} = 0.015$	
$\sigma_{\hat{p}} \approx 0.1225$	

The expected value of each probability distribution of sample proportions is the same as the population proportion, regardless of the sample size; however, the variance and standard deviation values change with the sample size. The variance decreases as n increases, indicating that as n increases, the probability distribution is packed more closely around the population proportion. The spread decreases while remaining centered on the population proportion. Let us construct the probability distributions of sample means and ranges to see if similar patterns emerge.

Table 5.1.16 Probability Distribution of Sample Means $n = 1, 2, 3$, and 4

	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$\mu_{\bar{x}}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$
$\sigma^2_{\bar{x}}$	0.24	0.09	0.04	0.015
$\sigma_{\bar{x}}$	≈ 0.4899	0.3	0.2	≈ 0.1225

Again, we encourage students struggling to understand these probability distributions and their construction to verify the results above by first building each probability distribution and then computing each measure. Click to check your work.

Probability distribution of sample means $n = 1$

Table 5.1.17 Probability Distribution of Sample Means $n = 1$ (5 Samples)

$n = 1$ **(5 Samples)**	
$\bar{x}$	$P(\bar{x})$
5	$\frac{1}{5}$
12	$\frac{1}{5}$

$n = 1$ (5 Samples)	
$\bar{x}$	$P(\bar{x})$
15	$\frac{1}{5}$
20	$\frac{1}{5}$
30	$\frac{1}{5}$
$\mu_{\bar{x}} = 16.4$	
$\sigma^2_{\bar{x}} = 69.84$	
$\sigma_{\bar{x}} \approx 8.3570$	

Probability distribution of sample means $n = 2$

Table 5.1.18 Probability Distribution of Sample Means $n = 2$ (10 Samples)

$n = 2$ (10 Samples)	
$\bar{x}$	$P(\bar{x})$
8.5	$\frac{1}{10}$
10	$\frac{1}{10}$
12.5	$\frac{1}{10}$
13.5	$\frac{1}{10}$
16	$\frac{1}{10}$
17.5	$\frac{2}{10}$,
21	$\frac{1}{10}$
22.5	$\frac{1}{10}$
25	$\frac{1}{10}$
$\mu_{\bar{x}} = 16.4$	
$\sigma^2_{\bar{x}} = 26.19$	
$\sigma_{\bar{x}} \approx 5.1176$	

Probability distribution of sample means $n = 3$

Table 5.1.19 Probability Distribution of Sample Means $n = 3)(\backslash($10 Samples)

$n = 3$ (10 Samples)	
$\bar{x}$	$P(\bar{x})$
$10.\bar{6}$	$\frac{1}{10}$
$12.\bar{3}$	$\frac{1}{10}$
$13.\bar{3}$	$\frac{1}{10}$

$n = 3$ (10 **Samples**)	
$\bar{x}$	$P(\bar{x})$
$15.\bar{6}$	$\frac{2}{10}$
$16.\bar{6}$	$\frac{1}{10}$
$18.\bar{3}$	$\frac{1}{10}$
19	$\frac{1}{10}$
$20.\bar{6}$	$\frac{1}{10}$
$21.\bar{6}$	$\frac{1}{10}$
$\mu_{\bar{x}} = 16.4$	
$\sigma^2_{\bar{x}} = 11.64$	
$\sigma_{\bar{x}} \approx 3.4117$	

Probability distribution of sample means $n = 4$

Table 5.1.20 Probability Distribution of Sample Means $n = 4$ (5 Samples)

$n = 4$ (5 **Samples**)	
$\bar{x}$	$P(\bar{x})$
13	$\frac{1}{5}$
13.5	$\frac{1}{5}$
16.75	$\frac{1}{5}$
17.5	$\frac{1}{5}$
19.25	$\frac{1}{5}$
$\mu_{\bar{x}} = 16.4$	
$\sigma^2_{\bar{x}} = 4.365$	
$\sigma_{\bar{x}} \approx 2.0893$	

We notice a similar trend with the probability distributions of sample means. Regardless of our sample size n, the expected value is the population mean. As n increases, the spread of our distributions decreases. We now look at a final example: sample ranges. Range, indeed all measures of spread, are not very informative (nor well-defined) if there is only one observation. Think about why that is. We will only consider samples with at least 2 observations.

Table 5.1.21: Probability Distribution of Sample Ranges $n = 1, 2, 3$, and 4

	$n = 2$	$n = 3$	$n = 4$
μ_{range}	11.6	17.4	21.6
σ^2_{range}	40.04	33.44	18.24
σ_{range}	≈ 6.3277	≈ 5.7827	≈ 4.2708

Click to verify probability distribution construction as needed.

Probability distribution of sample ranges $n = 2$

Table 5.1.22 Probability Distribution of Sample Ranges $n = 2$ (10 Samples)

$n = 2$ (10 **Samples**)	
range	$P(\text{range})$
3	$\frac{1}{10}$
5	$\frac{1}{10}$
7	$\frac{1}{10}$
8	$\frac{1}{10}$
10	$\frac{2}{10}$
15	$\frac{2}{10}$
18	$\frac{1}{10}$
25	$\frac{1}{10}$
$\mu_{\text{range}} = 11.6$	
$\sigma^2_{\text{range}} = 40.04$	
$\sigma_{\text{range}} \approx 6.3277$	

Probability distribution of sample ranges $n = 3$

Table 5.1.23 Probability Distribution of Sample Ranges $n = 3$ (10 Samples)

$n = 3$ (10 **Samples**)	
range	$P(\text{range})$
8	$\frac{1}{10}$
10	$\frac{1}{10}$
15	$\frac{1}{10}$
18	$\frac{2}{10}$
25	$\frac{1}{10}$
$\mu_{\text{range}} = 17.4$	
$\sigma^2_{\text{range}} = 33.44$	
$\sigma_{\text{range}} \approx 5.7827$	

Probability distribution of sample ranges $n = 4$

Table 5.1.24 Probability Distribution of Sample Ranges $n = 4$ (5 Samples)

$n = 4$ (5 **Samples**)	
range	$P(\text{range})$

$n = 4$ (5 Samples)	
range	$P(\text{range})$
15	$\frac{1}{5}$
18	$\frac{1}{5}$
25	$\frac{3}{5}$
$\mu_{\text{range}} = 21.6$	
$\sigma^2_{\text{range}} = 18.24$	
$\sigma_{\text{range}} \approx 4.2708$	

Having seen that the expected value of the probability distribution of sample ranges was not the population range in the case of $n = 3$, it is not surprising to see a similar situation in the other two sample sizes. The expected value of the probability distribution equaling the population parameter is not typical among the various statistics we consider. However, as n increases, we see a decrease in the spread of our probability distributions.

Sampling Distributions of Sample Statistics

Let us review what we know so far. We have constructed probability distributions of sample statistics under simple random sampling by computing a particular sample statistic for every possible sample of a specific size, n, and then determining the probability that these values occur. For some sample statistics, the probability distribution is centered around the population parameter; that is, for some statistics, the expected value of the probability distribution is the associated population parameter. Finally, we noticed that the spread of the probability distribution decreases as the sample size increases for each of the statistics studied so far.

We now introduce the probability distribution that much of inferential statistics is built upon, the **sampling distribution of sample statistics**. A sampling distribution is similar in nature to the probability distributions that we have been building in this section, but with one fundamental difference: rather than sampling using simple random sampling, the sampling method is to select randomly n objects, one at a time, from the population with replacement. Note that the order of selection will matter. As such, when considering a population of size N, there are N possibilities for each random selection, indicating that there are N^n samples to consider instead of the ${}_NC_n$ as with simple random sampling.

One may question why a distribution constructed from sampling with replacement takes priority in inferential statistics when the probability distributions above seem much more intuitive and easy to construct (having less samples to consider). As it so happens, when populations are large enough compared to the sample size (we will discuss this more later), the probability distributions of sample statistics constructed from simple random sampling are approximated well using the sampling distribution of sample statistics.

What intuition we have built from our previous constructions transfer directly to sampling distributions. As the sample size increases, the spread of the sampling distribution decreases. This should appeal to our basic intuition that larger samples better represent the population than smaller samples. For certain statistics, the expected value of the sampling distribution is the population parameter. We call such sample statistics **unbiased estimators.** Most introductory statistics books examine three unbiased estimators: sample means, proportions, and variances. The sample range is a biased estimator of the population range for the same reasons discussed above.

? Text Exercise 5.1.4

1. Within the context of the family of five from above and using a sample size of 3, construct the sampling distribution of sample proportions for the proportion of family members who wear glasses. Recall that the sampling method used in producing sampling distributions is selecting a member at random n times with replacement which means that order matters and the same member could be in the sample multiple times.

Answer

We begin our solution by counting the number of samples that we need to consider. Since the population size is $N = 5$ and our sample size is $n = 3$, we have $5^3 = 125$ different possible samples. That is a lot!

AAA	BAA	CAA	DAA	EAA
AAB	BAB	CAB	DAB	EAB
AAC	BAC	CAC	DAC	EAC
AAD	BAD	CAD	DAD	EAD
AAE	BAE	CAE	DAE	EAE
ABA	BBA	CBA	DBA	EBA
ABB	BBB	CBB	DBB	EBB
ABC	BBC	CBC	DBC	EBC
ABD	BBD	CBD	DBD	EBD
ABE	BBE	CBE	DBE	EBE
ACA	BCA	CCA	DCA	ECA
ACB	BCB	CCB	DCB	ECB
ACC	BCC	CCC	DCC	ECC
ACD	BCD	CCD	DCD	ECD
ACE	BCE	CCE	DCE	ECE
ADA	BDA	CDA	DDA	EDA
ADB	BDB	CDB	DDB	EDB
ADC	BDC	CDC	DDC	EDC
ADD	BDD	CDD	DDD	EDD
ADE	BDE	CDE	DDE	EDE
AEA	BEA	CEA	DEA	EEA
AEB	BEB	CEB	DEB	EEB
AEC	BEC	CEC	DEC	EEC
AED	BED	CED	DED	EED
AEE	BEE	CEE	DEE	EEE

We leave the computation of each sample proportion to the reader and tabulate the results in the following table.

Table 5.1.25 Sampling distribution of sample proportions

$\hat{p}$	$P(\hat{p})$
0	$\frac{8}{125}$
$\frac{1}{3}$	$\frac{36}{125}$
$\frac{2}{3}$	$\frac{54}{125}$
1	$\frac{27}{125}$

2. Compute the expected value, variance, and standard deviation of the sampling distribution of sample proportions found in the previous portion of this text exercise.

Answer

Table 5.1.26 Table of computations

$\hat{p}$	$P(\hat{p})$	$\hat{p} \cdot P(\hat{p})$	$\left(\hat{p} - \mu_{\hat{p}}\right)^2 \cdot P(\hat{p})$
0	$\frac{8}{125}$	0	$\left(0 - \frac{3}{5}\right)^2 \cdot \frac{8}{125} = \frac{72}{3125}$
$\frac{1}{3}$	$\frac{36}{125}$	$\frac{1}{3} \cdot \frac{36}{125} = \frac{12}{125}$	$\left(\frac{1}{3} - \frac{3}{5}\right)^2 \cdot \frac{36}{125} = \frac{576}{28125} = \frac{64}{312}$

$\hat{p}$	$P(\hat{p})$	$\hat{p} \cdot P(\hat{p})$	$\left(\hat{p} - \mu_{\hat{p}}\right)^2 \cdot P(\hat{p})$
$\frac{2}{3}$	$\frac{54}{125}$	$\frac{2}{3} \cdot \frac{54}{125} = \frac{36}{125}$	$\left(\frac{2}{3} - \frac{3}{5}\right))^2 \cdot \frac{54}{125} = \frac{54}{28125} = \frac{6}{312}$
1	$\frac{27}{125}$	$1 \cdot \frac{27}{125} = \frac{27}{125}$	$\left(1 - \frac{3}{5}\right))^2 \cdot \frac{27}{125} = \frac{108}{3125}$
$\mu_{\hat{p}} = \mathrm{E}(\hat{p}) = \frac{12}{125} + \frac{36}{125} + \frac{27}{125} = \frac{75}{125} = \frac{3}{5}$			
$\sigma^2_{\hat{p}} = \mathrm{Var}(\hat{p}) = \frac{72}{3125} + \frac{64}{3125} + \frac{6}{3125} + \frac{108}{3125} = \frac{250}{3125} = \frac{2}{25}$			
$\sigma_{\hat{p}} = \sqrt{\mathrm{Var}(\hat{p})} = \sqrt{\frac{2}{25}} \approx 0.2828$			

Sampling Distributions and Large Populations

The number of possible samples is quite large when we study most populations. As an example, FHSU enrolls about 3600 on-campus students in a typical semester. We might consider randomly sampling 30 students. In which case, the number of possible samples is about $4.8874 \cdot 10^{106}$. We naturally expect many possible sample statistic values with so many different possible samples. As such, we expect that sampling distributions take on large numbers of values and can be reasonably represented as continuous random variables, which can be understood using histograms and probability density functions.

With larger populations, we can no longer expect to construct all possible samples of a given size, n, from a population to develop an understanding of sampling distributions. However, that does not mean we cannot build a reasonably accurate representation. We can approximate sampling distributions by randomly sampling from all the possible samples and then constructing histograms to visualize the shape of the distribution. We build relative frequency histograms to estimate the probability distribution of the sampling distribution. This process is tedious but can be easily implemented with a computer.

? Text Exercise 5.1.5

The Online StatBook Project provides a program that operates with frequency counts rather than relative frequency counts. Since both preserve the general shape of a distribution, we can build an intuition about sampling distributions. Open the link and read the instructions page carefully. Then click on the "Begin" button at the top left part of your screen. Note: within this program, N stands for sample size, not population size.

1. Use the following settings for this program. **Parent Population:** Normal. **First Sampling Distribution:** Mean $N = 5$. **Second Sampling Distribution:** None $N = 5$. Click the "Animated" button and watch the animation.
   a. What do the boxes in the second distribution represent?
   b. What does the singular box in the third distribution mean?
   c. Construct a sampling distribution using 125 random samples (Reps=125), then $10, 125$ and $50, 125$ What is happening as we randomly sample more and more?

Answer

The five boxes in the second distribution represent the five observations randomly selected to form our single sample of size 5. The singular box in the third distribution is the sample mean of the sample in the second distribution. The distributions should look different since we are running a simulation randomly selecting samples, but they should be reasonably close to the figure provided below.

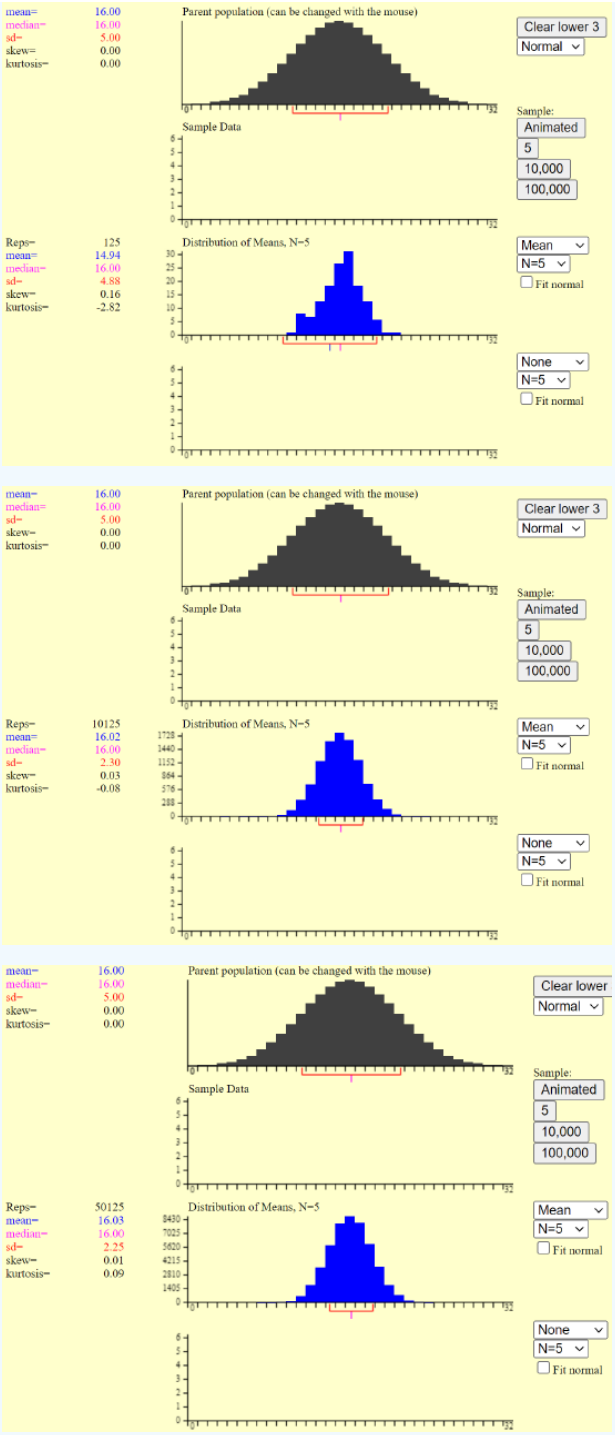


Figure 5.1.1: Sampling distribution simulation

As we take more and more samples of size 5 from the population, we eventually converge to a consistent shape. Once the distribution does not change much, we have a decent idea of the sampling distribution.

2. Use the following settings for this program. **Parent Population:** Normal. **First Sampling Distribution:** Mean $N = 2$ Fit Normal Checked. **Second Sampling Distribution:** Mean $N = 5$ Fit Normal Checked. Run the simulation using $100,000$ random samples to estimate the sampling distributions of sample means. Compare the two sampling distributions.

Answer

Once again, our distributions will not be the same as those produced by the simulation for you, but they should be quite similar.

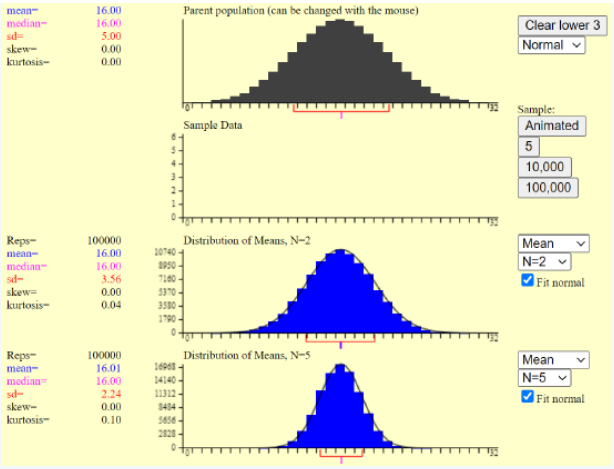


Figure 5.1.2 Sampling distribution simulation

Both sampling distributions are centered reasonably close to the population mean. When the sample size is 5, the sampling distribution is less spread out compared to the sampling distribution of sample size 2. Both sampling distributions have less spread than the parent population and fit the normal curve well.

3. Use the following settings for this program. **Parent Population:** Skewed. **First Sampling Distribution:** Mean $N = 2$ Fit Normal Checked. **Second Sampling Distribution:** Mean $N = 20$ Fit Normal Checked. Run the simulation using $100,000$ random samples to estimate the sampling distributions of sample means. Compare the two sampling distributions.

Answer

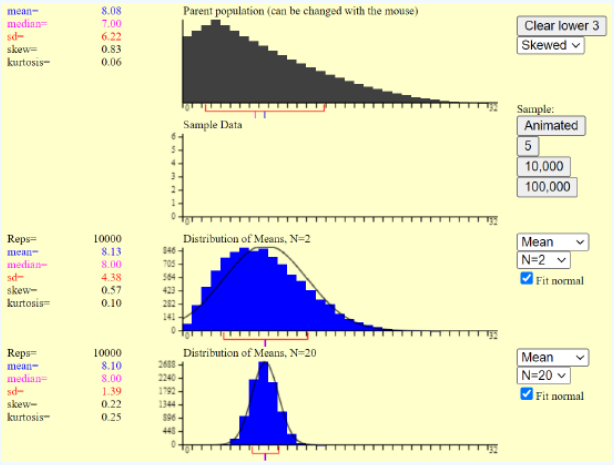


Figure 5.1.3 Sampling distribution simulation

Both sampling distributions are centered reasonably close to the population mean. The spread again decreases as the sample size increases. The first sampling distribution appears skewed to the right, just like the parent population, and is not normal. With a greater sample size, the second sampling distribution fits a normal curve much better.

4. Use the following settings for this program. **Parent Population:** Custom. **First Sampling Distribution:** Mean $N = 10$ Fit Normal Checked. **Second Sampling Distribution:** Mean $N = 25$ Fit Normal Checked. Using your mouse, construct a parent population so that when you run the simulation using $100,000$ random samples to estimate the sampling distributions of sample means, it does not appear to be normal.

Answer

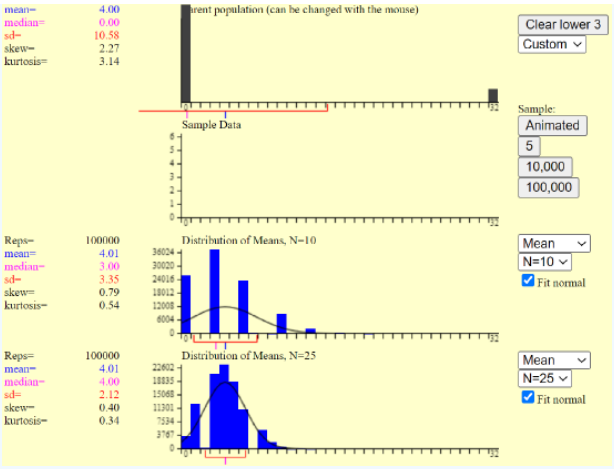


Figure 5.1.4 Sampling distribution simulation

It takes an extreme parent population for the sampling distributions not to fit a normal curve with a sample size of 25.

5. Use the following settings for this program. **Parent Population:** Skewed. **First Sampling Distribution:** Var (U) $N = 2$ Fit Normal Checked. **Second Sampling Distribution:** Var (U) $N = 25$ Fit Normal Checked. Run the simulation using $100,000$ random samples to estimate the sampling distributions of sample variances. Compare the two sampling distributions.

Answer

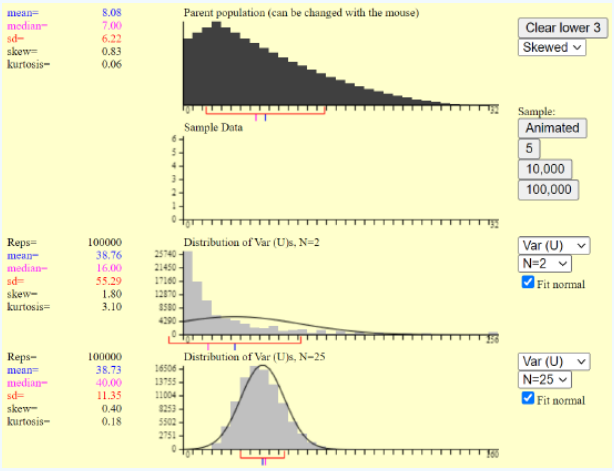


Figure 5.1.5 Sampling distribution simulation

Notice that the variance of the parent population $6.22^2 = 38.6884$is very close to the expected values (mean) of the sampling distributions. The sampling distributions are roughly centered on the population parameter. Recall that variance was the third unbiased estimator. Again, the spread of the sampling distributions decreases as the sample size increases. These sampling distributions, however, do not appear to be normally distributed. It seems closer when the sample size is 25, but it is still not a great fit.

6. In the previous parts of this text exercise, the mean of the sampling distributions has been very close to the population parameter. We do want to provide an intuition about unbiased estimators. Use the following settings for this program. **Parent Population:** Skewed. **First Sampling Distribution:** Range $N = 2, 5, 10, 16, 20, 25$ **Second Sampling Distribution:** None $N = 5$. For each sample size, run the simulation using $100,000$random samples to estimate the sampling distributions of sample ranges. Describe what happens to the expected value of the sampling distribution of sample ranges (the mean of the second distribution) as the sample size increases. How is this different from all other sampling distributions within this text exercise?

Answer

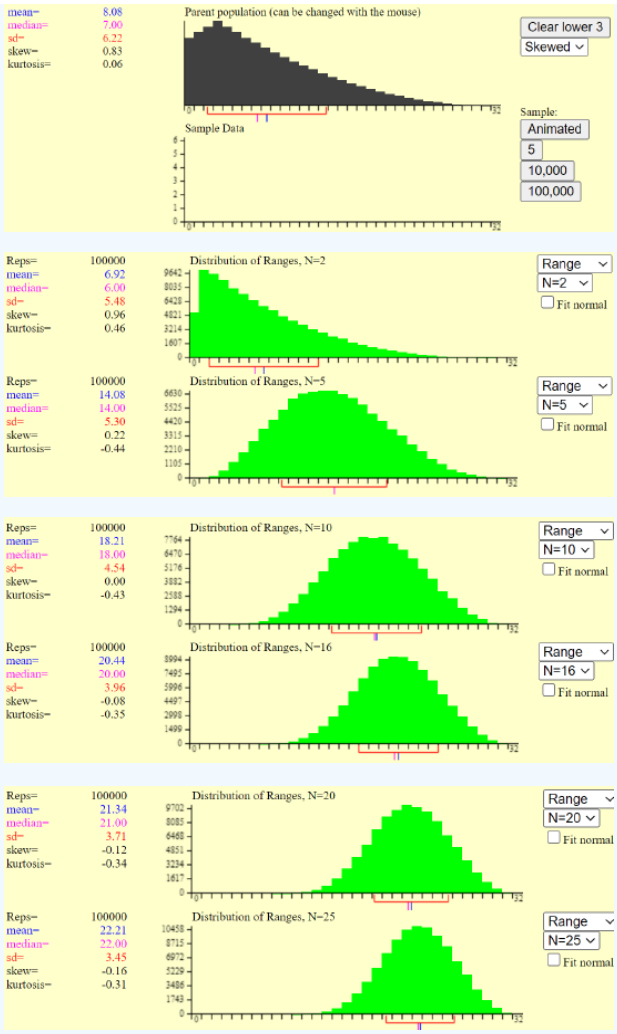


Figure 5.1.6 Sampling distribution simulation

The population range is about 28. If we look at the expected values of the sampling distributions of sample ranges as the sample size increases, we see that the expected values increase each time. This is in contrast to what happened with the sampling distributions of sample means and variances. Regardless of the sample size, the sampling distributions of sample means and variances had expected values close to the population parameter. The distributions were centered on the population parameter. Some sample means were less than the population mean; some were more. There was no inherent bias in the estimation using sample means or sample variances. This is not the case for sample ranges. We are guaranteed that the sample range is less than or equal to the population range. Why do you think this is true? Because of this, the sampling distribution cannot be centered on the population range unless we have a trivial population, making the sample range a biased estimator.

Bridging Theory and Application

While simple random sampling and random selection with replacement are two fundamentally different approaches to sampling, when populations are large enough and the sample size is not too large relative to the population size, we consider the two methods approximately interchangable in regards to the probability distributions of sample statistics that are produced. That is, when the size conditions are met (discussed in future sections), we utilize the sampling distribution of sample statistics for a given sample size n to understand the probability of sample statistic values from simple random samples of that same size n. It is important to note that statistical analyses have been developed for cases where the size conditions are not met, but that these considerations are beyond the scope of this text. Interested readers are encouraged to continue taking more advanced statistics courses.

- **9.1: Introduction to Sampling Distributions** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

5.2: Sampling Distribution of Sample Means

Learning Objectives

- Motivate, state, and apply the Central Limit Theorem (CLT)
- State the expected value (mean) and standard deviation of the sampling distribution of sample means
- Establish guides regarding sufficiently large sample sizes

The Utility of Sampling Distributions

To construct a sampling distribution, we must consider all possible samples of a particular size, n, from a given population. In reality, this is more complicated than studying the entire population since considering every possible sample requires studying every member of the population. If we have all the population data, why mess with all the samples? It is a valid question. The truth is that, in practice, statisticians do not construct sampling distributions by brute force; instead, they deduce key properties of the distribution. Inferential statistics are used to learn about a population by studying a sample, a subset of the population, not the entire population itself.

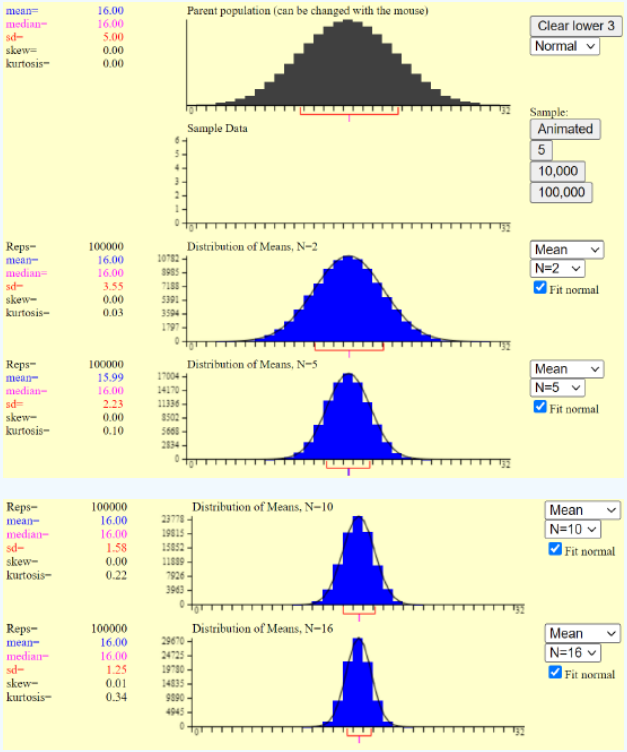
The Sampling Distribution of Sample Means

Using the computer simulation from the last section, we will consider the progression of sampling distributions of sample means from several populations as the sample size increases. Our previous work shows that the sampling distribution of sample means will be centered on the population mean and that the spread will decrease as the sample size increases. What can we say about the general shape of the sampling distributions of sample means regardless of the parent population?

? Text Exercise 5.2.1

The parent population (the distribution in black) is centered above 6 sampling distributions of sample means (the distributions in blue), starting with a sample size of 2 and ending with a sample size of 25. A normal curve has been fit to each of the sampling distributions. Which sampling distributions seem to fit the normal curve better? What trend do you notice across parent populations?

Parent Population: Normal



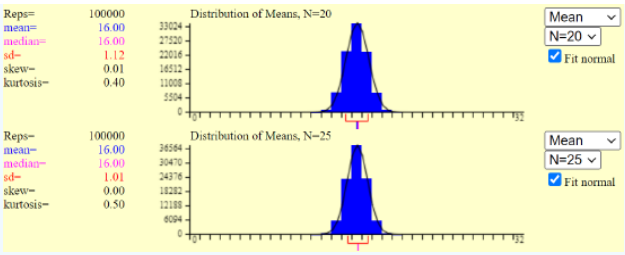


Figure 5.2.1: Sampling distributions of sample means for various sample sizes taken from a normal population

Parent Population: Uniform

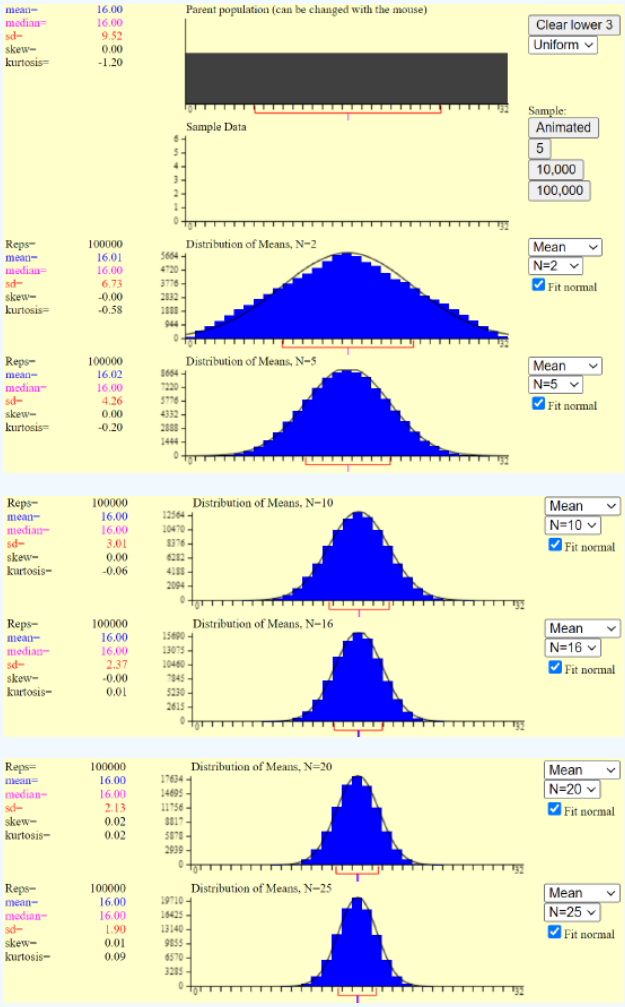


Figure 5.2.2 Sampling distributions of sample means for various sample sizes taken from a uniformly distributed population

Parent Population: Skewed

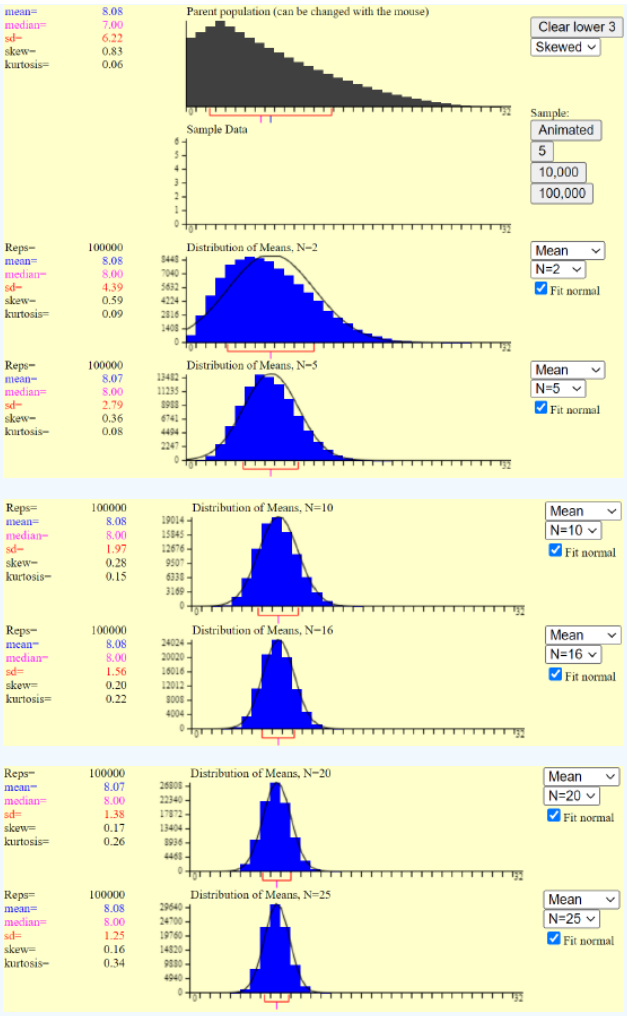


Figure 5.2.3 Sampling distributions of sample means for various sample sizes taken from a skewed population

Answer

Each sampling distribution from the normal parent population fits the normal curve well. All of the sampling distributions except the first sampling distribution with a sample size of 2 from the uniform parent population also fit the normal curve well. For the skewed parent population, it was not until the sample size reached 16 or 20 that the normal curve fit well. We see a trend that the sampling distributions of sample means eventually appear normal regardless of the parent distribution. For some parent distributions, larger sample sizes were necessary for the sampling distribution of sample means to appear to fit the normal curve.

Hopefully, we understand that the sampling distribution of sample means and the normal distribution are connected; furthermore, the sample size used to construct the sampling distribution also plays a role. If not, that is okay. We continue to learn and develop the relationship more formally. It is quite a remarkable result.

Note: Assessing Normality

Up to this point, we have assessed how well a distribution fits a normal curve visually. This level of discussion serves our purposes in an introductory statistics textbook. However, we want to alert the reader that there are analytical methods of assessing how well a normal curve fits a distribution. This process is commonly called assessing normality and is essential to serious statistical study and work.

Central Limit Theorem (CLT)

Given any infinite population with population mean μ and non-zero population variance of σ^2, as the sample size, n, increases, the sampling distribution of sample means approaches a normal distribution with mean $\mu_{\bar{x}} = \mu$ and variance $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$.

Recall that standard deviation is the square root of variance. We can also assert the normal distribution that the sampling distribution of sample means approaches as n increases has a standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. We will utilize this formula quite frequently.

Note: Infinite Populations

Notice that the Central Limit Theorem says, "given any infinite population." We have framed statistical inquiry in terms of understanding the world and people around us. At any given time, there are only finitely many humans, animals, or even atoms existing in our world. So, the statement of the Central Limit Theorem seems to exclude all these populations in which we may have interest, but hope is not lost.

The Central Limit Theorem, as stated above, is a beautiful work of mathematical and statistical theory. There is often a gap between theory and practice that can be bridged satisfactorily. We are in such a case, and our bridge is the notion of a practically infinite population relative to the sample size in consideration. A population may be understood as practically infinite if the sample size of interest is less than 5% of the population size (recall how we saw this earlier with the assumption of independence). The Central Limit Theorem holds in practice for practically infinite populations. Indeed, this rule of thumb of 5% also serves as our threshold at which we treat simple random sampling and sampling with replacement as interchangeable regarding the probability distributions of sample statistics.

Let us think about what the Central Limit Theorem is saying. It claims that if you pick any population (regardless of shape) and look at all samples of size n (for n sufficiently large), their means will be (approximately) normally distributed. We could start with any population, even the craziest of shapes, and an orderly bell curve emerges from the chaos of random selection. This should come as a surprise to someone hearing this for the first time. One would expect that nothing can be said about the sampling distribution; if the population shape is chaotic and we are selecting samples from it at random, then we would expect the sampling distribution to be chaotic as well. After all, the sampling distributions of the range, median, mode, and many other statistics do not follow such nice behavior in general. They behave exactly as the default intuition would expect: chaotically and unpredictably. However, there is something special about the mean. We will see later that the mean is not the only statistic to exhibit nice behavior.

The Central Limit Theorem is the reason the field of inferential statistics exists. The fact that we always get a normal distribution enables us to answer questions in inferential statistics intelligently and precisely, as we shall see now.

Applying the Central Limit Theorem

The Central Limit Theorem clearly states the ideas we have been exploring over the last two sections.

- The sampling distribution of sample means has an expected value (mean), the population mean.
- The spread of the sampling distribution of sample means decreases as n increases because $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The population standard deviation, σ, is fixed; so, as n increases, we have a fixed number divided by larger and larger numbers making the quotient smaller.
- Finally, the sampling distribution of sample means gets closer and closer to the normal curve as n increases.

As we have seen, the rate at which the sampling distribution's shape becomes normal differs based on the parent population. If the parent population is normal, every sampling distribution appears approximately normal. In the other cases, we needed the sample size to be larger. How can we know if a given sample size is large enough to say that the sampling distribution of sample means is approximately normal? We now provide, without proof, the current knowledge and standards of the statistical community in this regard.

If the parent population is normally distributed, the sampling distribution of sample means will be normally distributed for every sample size, n.

Statisticians have long agreed that for many of the distributions commonly found in medicine, the social sciences, and the natural sciences, a sample size larger than 30 would produce a sampling distribution of sample means that is approximately normal.

Our statistical research may find a population in which a smaller number produces an approximately normal sampling distribution of sample means. We would not know this when we first began studying the population. On the other hand, we may find a population in which a much larger sample would be necessary to produce an approximately normal sampling distribution of sample means. Such distributions are being studied in economics and finance. How can we feel confident in our practice of statistics, especially if we conduct statistical research as part of our profession?

The parent distributions that require larger sample sizes to obtain approximately normal sampling distributions of sample means are most likely extremely skewed or have outliers. If there is no such intuition, we recommend using an initial sample size larger than 30 and testing the data for the presence of outliers or extreme skew (a histogram will probably suffice). If either outliers or extreme skew are detected, proceed by increasing the sample size and repeating the data collection process.

? Text Exercise 5.2.2

The grade distribution for a particular instructor's statistics course (over many years with thousands of students) is negatively skewed with a mean of 71% and a standard deviation of 20%. Compute the probability that the average of a random sample for the indicated sample size is within 2 percentage points of the population mean. What do you notice about the probabilities as n increases?

1. $n = 36$

Answer

Since the random sample is larger than 30, the sampling distribution of sample means is approximately normal with a mean $\mu_{\bar{x}} = 71$ percent and a standard deviation $\sigma_{\bar{x}} = \frac{20}{\sqrt{36}}$ percent. We are determining the probability that the average of the sample is within 2 percentage points of 71 percent: $P(69 < \bar{x} < 73)$. Sketch a picture and compute the probability with technology.

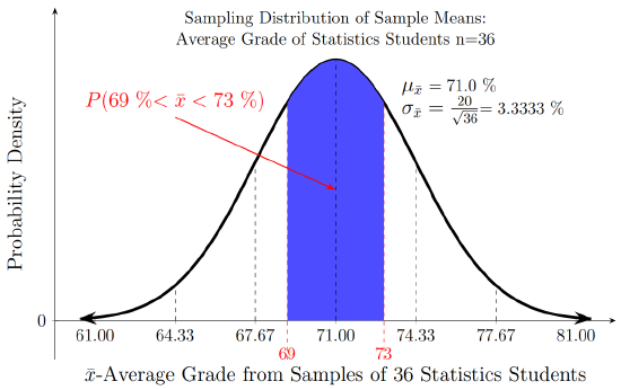


Figure 5.2.4 Sampling distribution of sample means

$$P(69 < \bar{x} < 73) = \text{NORM.DIST}(73, 71, \frac{20}{\sqrt{36}}, 1) - \text{NORM.DIST}(69, 71, \frac{20}{\sqrt{36}}, 1) \approx 45.1494\%$$

2. $n = 72$

Answer

The problem setup remains the same; we update the sample size to 72.

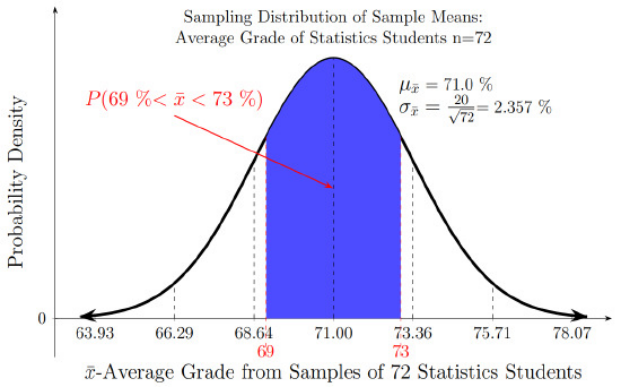


Figure 5.2.5 Sampling distribution of sample means

$$P(69 < \bar{x} < 73) = \text{NORM.DIST}(73, 71, \frac{20}{\sqrt{72}}, 1) - \text{NORM.DIST}(69, 71, \frac{20}{\sqrt{72}}, 1) \approx 60.3856\%$$

3. $n = 144$

Answer

The problem setup remains the same; we update the sample size to 144.

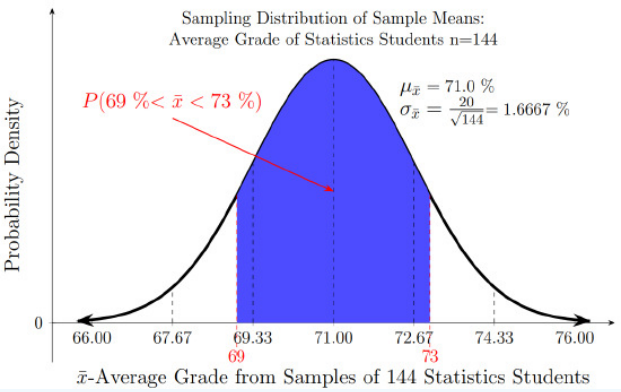


Figure 5.2.6 Sampling distribution of sample means

$$P(69 < \bar{x} < 73) = \text{NORM.DIST}(73, 71, \frac{20}{\sqrt{144}}, 1) - \text{NORM.DIST}(69, 71, \frac{20}{\sqrt{144}}, 1) \approx 76.9861\%$$

4. $n = 288$

Answer

The problem setup remains the same; we update the sample size to 288.

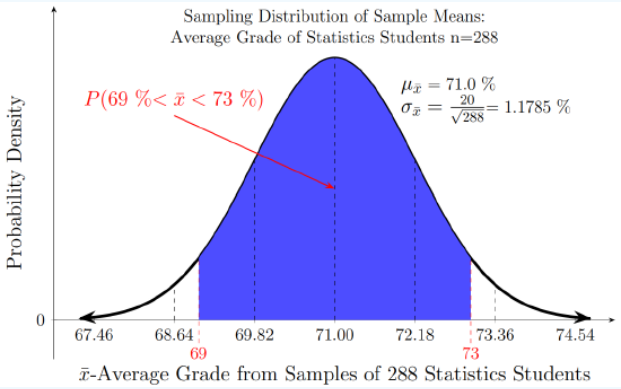


Figure 5.2.7 Sampling distribution of sample means

$$P(69 < \bar{x} < 73) = \text{NORM.DIST}(73, 71, \frac{20}{\sqrt{288}}, 1) - \text{NORM.DIST}(69, 71, \frac{20}{\sqrt{288}}, 1) \approx 91.0314\%$$

As the sample size increases, the standard deviation of the sampling distribution decreases. The interval from 69 to 73 encompasses more standard deviations around the mean, and the probability of the sample mean falling in that interval increases. If n were to increase to the size of the population, we would see a 100% chance of the sample mean being within 2 points of the actual mean. The larger n, the more likely the sample mean is close to the population mean.

Text Exercise 5.2.3

The heights of adult females are normally distributed with a mean of 64 inches and a standard deviation of 2.5 inches.

1. Determine the probability of randomly selecting four adult females whose average height is less than 5 feet 2 inches.

Answer

We randomly selected four adult females from the population and considered their average height. We took a sample of size $n = 4$ and considered the average height, $\bar{x}$, of the sample. We are interested in the following probability: $P(\bar{x} < 62)$. Note that we want to use the same units throughout 5 feet 2 inches is $5 \cdot 12 + 2 = 62$ inches. We must turn to the sampling distribution of sample means to answer the probability question. To compute probabilities, we must know what the probability distribution is. Constructing the probability distribution is out of the question here. We want to utilize the Central Limit Theorem (CLT). When considering the CLT, we ensure our sample size is large enough to assert that the sampling distribution of sample means is approximately normal. Usually, this means we want $n > 30$. This, however, is not the case in this scenario as $n = 4$. To proceed in this scenario, we note that the problem states that the heights of adult females are normally distributed. If the parent population is normally distributed, so are all the sampling distributions of sample means. We know that the sampling distribution of sample means is normally distributed with a mean $\mu_{\bar{x}} = \mu = 64$ inches and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{4}} = 1.25$ inches. We sketch a picture and compute the probability using technology.

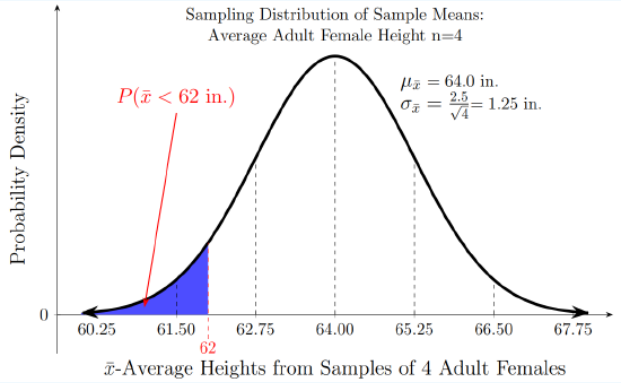


Figure 5.2.8 Sampling distribution of sample means

$$P(\bar{x} < 62) = \text{NORM.DIST}(62, 64, 1.25, 1) \approx 5.4799\%$$

2. Determine the probability of randomly selecting two adult females with an average height within 3 inches of the population mean.

Answer

We are in the context of randomly selecting multiple adult females from the population and considering their average height. We are only sampling 2 adult females; so, $n = 2$. We utilize the fact that the parent population is normal and that the sampling distribution of sample means when $n = 2$ is normal with a mean $\mu_{\bar{x}} = \mu = 64$ inches and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{2}} \approx 1.7678$ inches. An average height is within 3 inches of the population mean if it is larger than $64 - 3 = 61$ inches and smaller than $64 + 3 = 67$ inches. So, we are interested in $P(61 < \bar{x} < 67)$. We sketch a picture and compute the probability using technology.

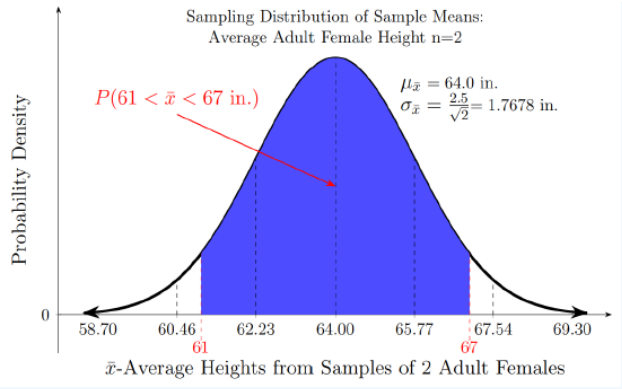


Figure 5.2.9 Sampling distribution of sample means

$$P(61 < \bar{x} < 67) = \text{NORM.DIST}(67, 64, \frac{2.5}{\sqrt{2}}, 1) - \text{NORM.DIST}(61, 64, \frac{2.5}{\sqrt{2}}, 1) \approx 0.955157 - 0.044843 \approx 91.0314\%$$

Text Exercise 5.2.4

Recall from Text Exercise 4.5.1 that the daily growth in the height of wheat plants during a particular stage of development is believed to be uniformly distributed between $\frac{1}{2} = 0.5$ and $\frac{5}{4} = 1.25$ inches and as such has a mean of $\frac{0.5+1.25}{2} = 0.875$ inches and the standard deviation is \(\frac{1.25-0.5}{\sqrt{12}}\) $= \frac{0.75}{\sqrt{12}}$ inches. Determine the probability of randomly selecting 48 wheat plants (during that particular stage of development) with an average daily growth that is greater than 0.9 inches.

Answer

We are randomly selecting 48 wheat plants from the population and considering their average daily growth. We are taking a sample of size $n = 48$ and considering the average height, $\bar{x}$, of the sample. We are interested in the following probability: $P(\bar{x} > 0.9)$. Since the sample size is greater than 30, we can apply the CLT to say that the sampling distribution of sample means is approximately normal with a mean $\mu_{\bar{x}} = \mu$ inches and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ inches. Thus, $\mu_{\bar{x}} = 0.875$ inches and $\sigma_{\bar{x}} = \frac{\frac{0.75}{\sqrt{12}}}{\sqrt{48}} = \frac{0.75}{24} = 0.03125$inches. Let us sketch a picture and compute the probability of interest.

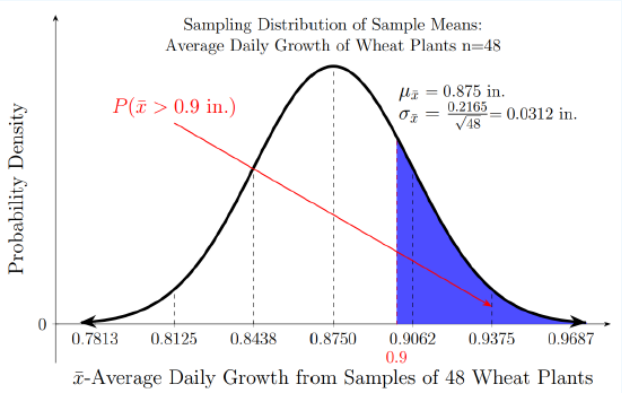


Figure 5.2.10 Sampling distribution of sample means

$$P(\bar{x} > 0.9) = 1 - \text{NORM.DIST}(0.9, 0.875, 0.03125, 1) \approx 1 - 0.788145\% \approx 21.1855\%$$

5.3: Sampling Distribution of Sample Proportions

Learning Objectives

- State the relationship between the sampling distribution of sample proportions($\hat{p}$) and a normal distribution.
- State the expected value (mean) and standard deviation of the sampling distribution of sample proportions.
- State the requirements for modeling the sampling distribution of sample proportions with a normal distribution.
- Apply the above to reasonably predict the proportion measures of various samples (all of the same size n) from a population.

Review and Preview

In regard to a random variable of a population, we have discussed the importance of understanding how various samples taken from the population produce different measures from each other as well as from the population's related measures. In the first section of this chapter, we saw that some statistical measures (such as samples' means, samples' variances, and samples' proportions for samples of a specific size n) are considered unbiased since the various samples' statistics tend to crowd around the actual population's parameter. However, there are other statistical measures (such as samples' ranges, samples' standard deviation, and samples' medians) that do not behave this way, and hence are considered biased estimators.

Digging deeper in the last section, we have seen how the sample means from all possible samples are actually very predictable as a group. Under certain requirements, the sample means for samples of one specific size, n, act as a random variable themselves and that, although we can't predict what will happen with any one chosen simple random sample, the collection of all simple random samples' means form a normal distribution called a sampling distribution. Furthermore, this sampling distribution's mean value is the same as the population's mean value and the spread (standard deviation) in the sampling distribution is smaller than the standard deviation of the population. In notational form, we designated this with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Now we embark on a similar investigation of the sampling distribution's behavior for the proportion measure. Recall from here in Section 5.1 that we have seen an example of building a sampling distribution for a small population (our family of five) in which our "proportion' variable of interest was the proportion of the family members that wear glasses. In this small example, there were three of the five family members that wear glasses making our population's proportion measure $p = \frac{3}{5}$; however, when we selected samples of size three from the population, the ten different samples produced various sample proportion measures $(\hat{p})$, none of which were the same as the population's parameter measure of $p = \frac{3}{5}$. However, the sampling distribution of these various sample proportions--the probability distribution of sample proportions--had a distribution mean that did match the population's proportion. That is, we saw in this small example that $\mu_{\hat{p}} = p$ even though $p \neq \hat{p}$ for any of the actual samples of size three. So, as in the general sampling distribution of sample means, we wonder if the sampling distribution of sample proportions is just as predictable.

The Sampling Distribution of Sample Proportions

First, we need to recognize that sample proportion measures fall into the realm of a binomial experiment with the number of trials being the sample size, n, and the probability of success, p, is the proportion of that population meeting the definition of "success" in the binomial experiment. Each time we select a member to be part of our sample, we are performing a binomial experiment. As a reminder from Section 4.3, recall that the random variable, X, of a binomial experiment was the number of successes that could occur with a sample of size n taken from the population and that the possible values for the random variable were $0, 1, 2, \ldots, n$. In general, it is possible to have a sample in which none (0) of the sample group met the "success" definition of the binomial experiment or a sample in which only 1 of the n was a "success," or that 2 of the n, ... , all the way to all n of the n were a success. We could build the binomial probability distribution from such information based on combination counts and our probability multiplication rule. We see this relationship established more formally below concerning a large population in which many, many samples of a size n are possible, or, in the case of a small finite population, with random sampling with replacement on samples of a size n being possible. Using our prior concepts of Section 4.3, we can build the binomial distribution concerning samples of size $n = 3$ coming from a population in which the population proportion measure of success is $p = \frac{3}{10} = 30\%$. Hence, the population proportion measure of failure is $q = \frac{7}{10} = 70\%$. Using our binomial distribution approach of Section 4.3, we produce the following distribution table.

Table 5.3.1: Binomial probability distribution

Number of Successes in n trials: X	Probability $P(x) = P(\hat{p})$
0	$_3C_0 \cdot \left(\frac{3}{10}\right)^0 \cdot \left(\frac{7}{10}\right)^3 = 1 \cdot (1) \cdot \left(\frac{343}{1000}\right) = \frac{343}{100} = 34.3\%$
1	$_3C_1 \cdot \left(\frac{3}{10}\right)^1 \cdot \left(\frac{7}{10}\right)^2 = 3 \cdot \left(\frac{3}{10}\right) \cdot \left(\frac{49}{100}\right) = \frac{441}{1000} = 44.1\%$
2	$_3C_2 \cdot \left(\frac{3}{10}\right)^2 \cdot \left(\frac{7}{10}\right)^1 = 3 \cdot \left(\frac{9}{100}\right) \cdot \left(\frac{7}{10}\right) = \frac{189}{1000} = 18.9\%$
3	$_3C_3 \cdot \left(\frac{3}{10}\right)^3 \cdot \left(\frac{7}{10}\right)^0 = 1 \cdot \left(\frac{27}{1000}\right) \cdot (1) = \frac{27}{1000} = 2.7\%$

We also recall that the binomial probability distribution was found to have an expected value or mean of $\mu = n \cdot p = 3 \cdot 0.30 = 0.90$ and variance of $\sigma^2 = n \cdot p \cdot q = 3 \cdot 0.30 \cdot 0.70 = 0.63$, and a standard deviation of $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{0.63} \approx 0.793725$. As the binomial distribution demonstrates and

our past work confirms, not all samples of size n taken from a population will have the same number of success x; and the related proportion of successes $\frac{x}{n}$ will tend to vary from sample to sample. We saw this in our small family member example mentioned in Section 5.1.

We now connect the sample proportion measures to this binomial distribution. Notice that our random variable X on the number of successes can be transformed into sample proportion measures simply by dividing by the sample size. For example, suppose we have a sample of size $n = 3$ with $x = 2$ successes. In that case, the sample's proportion measure of success is $\hat{p} = \frac{2}{3} \approx 66.7\%$. We can do this for each possible value of our random variable X, producing the following distribution table.

Table 5.3.2: Binomial probability distribution

Number of Successes in n trials: X	Proportion of Success $\hat{p} = \frac{x}{n}$	Probability $P(x) = P(\hat{p})$
0	$\frac{0}{3} = 0.00 = 0\%$	34.3%
1	$\frac{1}{3} \approx 0.333.00 = 33.3\%$	44.1%
2	$\frac{2}{3} \approx 0.667 = 66.7\%$	18.9%
3	$\frac{3}{3} = 1.00 = 100\%$	2.7%

This $\hat{p}$-probability distribution is the sampling distribution, and below is a graphic of that binomial distribution and its related sampling distribution of sample proportions.

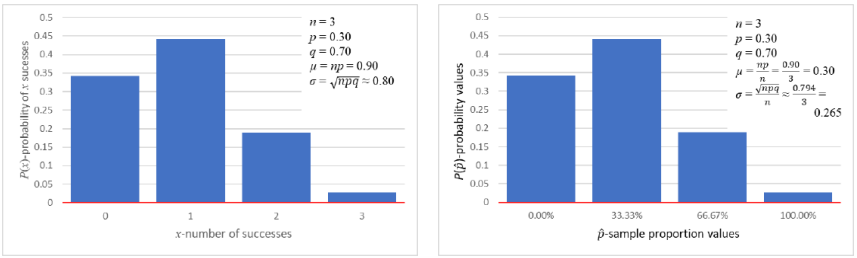


Figure 5.3.1: Binomial distribution (left) transformed into the sampling distribution of sample proportions (right)

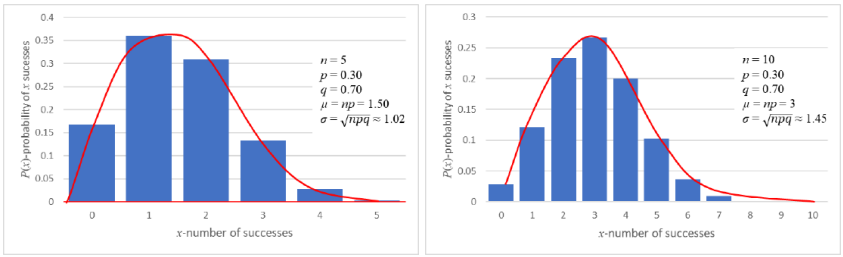
We should recognize that the only change occurring when moving from the binomial distribution to the sampling distribution is a rescaling of the horizontal axis, which results in a rescaling of the mean, μ, and the standard deviation, σ, both caused by our division by the sample size $n = 3$.

Of course, with this sampling distribution table or with its graph, we understand what values occur for the sample proportions concerning all the various simple random samples of size $n = 3$. For example, we may wish to know how probable it is to find a random sample of size 3 from this population in which the sample proportion is below 50%...that is to find $P(\hat{p} < 0.50)$. From our distribution, we can determine that

$$P(\hat{p} < 0.50) = P(\hat{p} = 0.00) + P(\hat{p} = 0.33) \approx 34.3\% + 44.1\% = 78.4\%.$$

That is, over 75% of random samples of size 3 from this population will produce sample proportion $\hat{p}$ measures below 50%.

The process from above is how we build the sampling distribution of sample proportions with **small** sample sizes, as the binomial distributions only have a few possible outcomes for the random variable "a number of successes." However, these binomial distribution tables get very large and cumbersome if working with large samples. For example, if dealing with samples of size $n = 100$, we would need to build a table with 101 rows with sample proportion measures from $0\%, 1\%, 2\%, \ldots, 100\%$.We can quickly see how even larger but often used sample sizes such as 2000 could be difficult to work with. To find a way around this, we continue our theory building in which the population proportion is $p = 0.30$, and then work with various samples of size $5, 10, 25$, and 50. Using the same approach as above, we can produce the binomial distribution tables (not shown), and then the graphics for each of those distribution tables, shown below in Figure 5.3.2. Notice how the binomial distributions become more bell-shaped and symmetrical as the sample size gets larger, specifically in the distributions for the largest two sample sizes of $n = 25$ and $n = 50$.



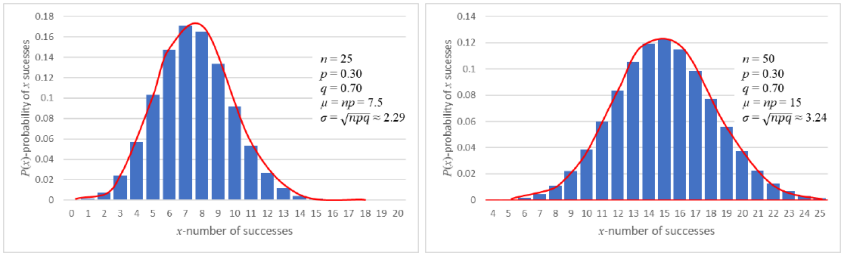


Figure 5.3.2: Binomial distributions approaching a normal distribution

This same behavior tends to occur regardless of the actual population proportion value, p, provided the sample size, n, is sufficiently large. Without delving too deeply into the underlying mathematical reasoning, the binomial distribution can be considered an approximately normal distribution in behavior provided $n \cdot p > 5$ and $n \cdot q > 5$. Statisticians wishing to be even more conservative to achieve greater accuracy in the use of a normal distribution to approximate a binomial distribution will often require $n \cdot p > 10$ and $n \cdot q > 10$. The above also demonstrates that as n increases, the normal distribution approximation becomes a better and better fit for the binomial distribution. We also have the mean and standard deviation of this approximating normal distribution due to our knowledge of the binomial distribution; that is, our normal distribution approximation to the binomial distribution will also have a mean of $\mu = n \cdot p$ and standard deviation of $\sigma = \sqrt{n \cdot p \cdot q}$.

Now let us examine how this is related to the "sample proportion" random variable instead of the "number of success" random variable. As in Figure 5.3.1, we can adjust to "proportion" measures instead of "number of success" in the distributions by dividing each of our random variable's x-values by the sample size n. As can be seen below in Figure 5.3.3 this change to the proportion variable only causes a change in the scaling of the x-axis and measures related to that axis, but does not change the distribution probability measures nor the basic shape of the distribution.

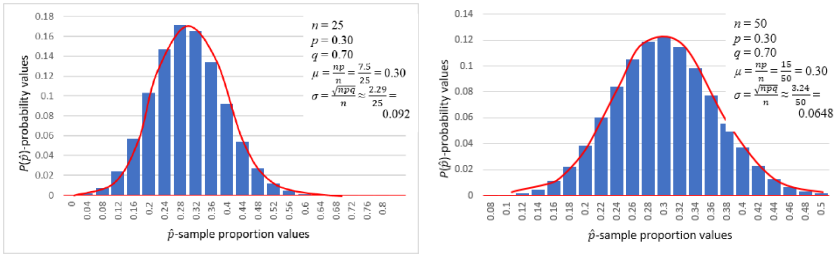


Figure 5.3.3: Binomial Distributions rescaled to Proportion Distributions

Therefore, we note that under sufficient requirements, our sampling distribution for the proportion values will be approximately a bell-shaped distribution with key measures of $\mu_{\hat{p}} = \frac{n \cdot p}{n} = p$ and $\sigma_{\hat{p}} = \frac{\sqrt{n \cdot p \cdot q}}{n} = \sqrt{\frac{n \cdot p \cdot q}{n^2}} = \sqrt{\frac{p \cdot q}{n}}$. We can then use a normal probability distribution to estimate the binomial probability values over intervals; a normal probability distribution with appropriate probability density function with the same mean and standard deviation as we found above. Similar to the Central Limit Theorem (CLT) for predicting the distribution of all possible sample means in a specific situation, we have a theorem for predicting the distribution of all possible sample proportions within a particular situation.

Sampling Distribution of Sample Proportions

Given a binomial situation within a population of interest in which the following conditions are known:

- the requirements for a binomial distribution are met with
  - the population proportion of interest (probability of success) is p
  - the complement proportion (probability of failure) is $q = 1 - p$
  - the sample size (number of finite trials) is n
- the requirements of $n \cdot p > 5$ and $n \cdot q > 5$ are met

then the sampling distribution of all possible sample proportions can be approximated as a normal random variable with $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}}$.

If desiring more reliable measures in using the normal distribution to approximate the distribution of sample proportions, we instead use the more conservative requirements of $n \cdot p > 10$ and $n \cdot q > 10$. If the values of $n \cdot p$ and $n \cdot q$ do not surpass at least 5, then we do not approximate with a normal distribution but instead use the binomial probability distribution adjusted to sample proportions to model the sampling distribution.

Applying the Theorem on Sampling Distribution of Sample Proportions

With a specific population of interest, our theorem allows us to understand which sample proportions are likely to happen and which are unlikely. This knowledge is important for understanding how we can have confidence in predicting a population's proportion from a single sample as we turn to inferential statistics in the next chapter. So, to prepare for this, we apply this theorem with the following text exercises.

Text Exercise 5.3.1

It is believed that 21% of all U.S. female adults are over 66 inches in height. Determine the probability of selecting a simple random sample of U.S. female adults in which over 30% of the sample group is over 66 inches tall for each sample size given below. What do you notice about the probabilities as n increases?

1. $n = 25$

Answer

We first note that this can be considered a "binomial" experiment in which we are defining "success" as a U.S. female adult having a height measure over 66 inches. The population's proportion is $p = 21\% = 0.21$ and the "failure" proportion is $q = 1 - 0.21 = 0.79$. Instead of building a binomial distribution to answer the question, we can answer this question using our above theorem since $n \cdot p = 25 \cdot 0.21 = 5.25$ and $n \cdot q = 25 \cdot 0.79 = 19.75$ are both values above 5. (We also note that 5.25 is not much above 5 and right on the border of meeting the requirements of the theorem; in general, when getting close in value to the requirements, we understand that our measures are not as reliable and should not be used for highly important or costly decision making.) Based on our developed theory above, the sampling distribution of sample proportions is approximately normal with a mean $\mu_{\hat{p}} = p = 0.21$ and a standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{0.21 \cdot 0.79}{25}}$. Sketching a graphic of this normal distribution, we see the following.

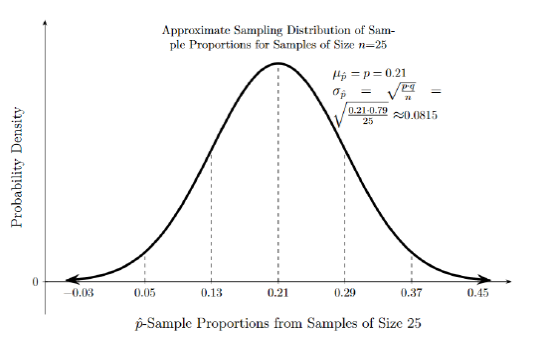


Figure 5.3.4 Approximate sampling distribution of sample proportions

We can compute the approximate probability for randomly selecting a sample of size 25 in which the proportion measure from that sample is larger than 30%; that is, we can find approximately from our normal distribution the value of $P(\hat{p} > 30\%)$ by finding the shaded region displayed below in our related sampling distribution.

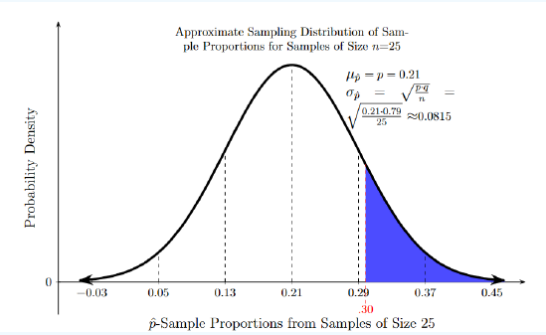


Figure 5.3.5 Approximate sampling distribution of sample proportions

Using our spreadsheet's NORM.DIST function concerning the normal distribution above, we produce:

$$P(\hat{p} > 30\%) = 1 - \text{NORM.DIST}(0.30, 0.21, \sqrt{\frac{0.21 \cdot 0.79}{25}}, 1)) \approx 13.462\%$$

About $13.462\%$of all possible samples of size $n = 25$ from the population of U.S. female adults will have over $30\%$of the women in the sample being over 66 inches tall. Stated equivalently, the probability of randomly selecting 25 U.S. female adults in which over 30% of the women are over 66 inches tall is about 13.462%.Although not highly likely, such a sample result would generally not be considered unusual.

2. $n = 50$

Answer

The problem setup remains the same; we update for the sample size of 50, noting that we still meet our requirements with $n \cdot p = 50 \cdot 0.21 = 10.5$ and $n \cdot q = 50 \cdot 0.79 = 39.5$ both above 10. In our graphic below, we point out for emphasis the scaling change that occurred in the

horizontal axis as compared to part 1. of this text exercise above.

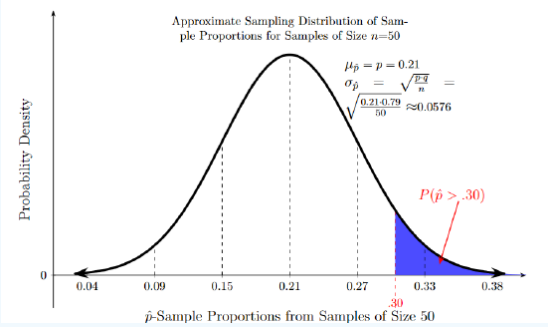


Figure 5.3.6 Approximate sampling distribution of sample proportions

$$P(\hat{p} > 30\%) = 1 - \text{NORM.DIST}(0.30, 0.21, \sqrt{\frac{0.21 \cdot 0.79}{50}}, 1)) \approx 5.9092\%$$

The probability of randomly selecting 50 U.S. female adults in which over 30% of the women samples are over 66 inches tall is about 5.9092%.We note that this is a less likely outcome as compared to such in samples of size 25.

3. $n = 100$

Answer

The problem setup remains the same; we update the sample size to 100 and note that we still meet our requirements with $n \cdot p = 100 \cdot 0.21 = 21.0$ and $n \cdot q = 100 \cdot 0.79 = 79.0$ both well above 10.

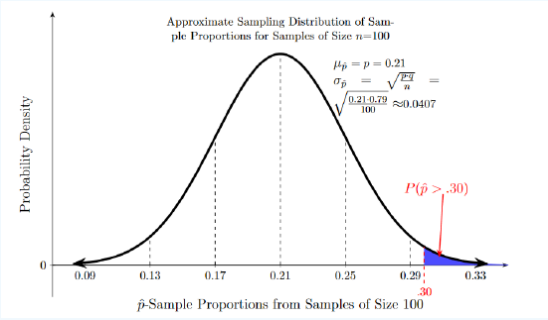


Figure 5.3.7 Approximate sampling distribution of sample proportions

$$P(\hat{p} > 30\%) = 1 - \text{NORM.DIST}(0.30, 0.21, \sqrt{\frac{0.21 \cdot 0.79}{100}}, 1)) \approx 1.3565\%$$

4. $n = 250$

Answer

We again update to the sample size of 250 and note that we easily meet our restrictive requirements with $n \cdot p = 250 \cdot 0.21 = 52.5$ and $n \cdot q = 250 \cdot 0.79 = 197.5$.

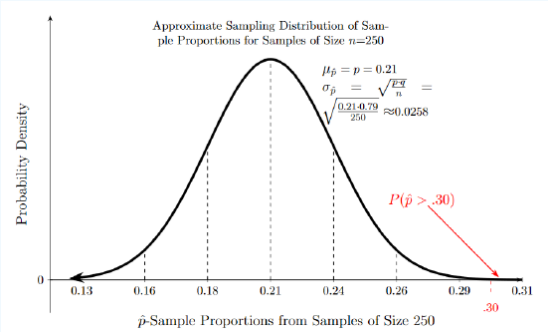


Figure 5.3.8 Approximate sampling distribution of sample proportions

$$P(\hat{p} > 30\%) = 1 - \text{NORM.DIST}(0.30, 0.21, \sqrt{\frac{0.21 \cdot 0.79}{250}}, 1)) \approx 0.0238\%$$

We note in this last case that it is extremely unlikely to randomly select a sample of 250 U.S. adult women in which over 30% of the sample group are over 66 inches tall.

Finally, looking across all four exercises, we notice that as the sample size increases, the standard deviation of the sampling distribution decreases. In paying attention to the horizontal axis scale as it changes through these exercises, if the sample size, n, were to increase close to the size of the population, we would see almost a 100% chance of the various sample proportions being very very close to the population's proportion of 21%. That is, the larger n is, the various possible random sample proportions will usually be very close to the population's proportion. In the next chapter, we will develop more specific measures for the vague term "close."

? Text Exercise 5.3.2

In Kansas, 35% of adults over 25 years old have a bachelor's degree or higher.

1. Determine the probability of randomly selecting fifty Kansas adults over 25 years old in which less than 20% of the sample group have a bachelor's degree or higher.

Answer

We are imagining randomly selecting fifty individuals from the population of Kansas adults over 25 years of age and determining the sample proportion $(\hat{p})$ of those selected who have a bachelor's degree or higher (hence a binomial situation). We should understand at this point that different samples will produce different $\hat{p}$ values, and in using random sampling, we do not know which sample we will get. However, we are interested in the following probability: $P(\hat{p} < 20\%)$. So, we must turn to the sampling distribution of sample proportions to answer the probability question. Utilizing our developed theory and the given information, we note that $n = 50$, $p = 35\% = 0.35$, and $q = 65\% = 0.65$. Since $n \cdot p = 50 \cdot 0.35 = 17.5$ and $n \cdot q = 50 \cdot 0.65 = 32.5$ are both larger than 5, we can reasonably approximate the sampling distribution of the various possible $\hat{p}$ values as a normal distribution with a mean $\mu_{\hat{p}} = p = 0.35$ and a standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{0.35 \cdot 0.65}{50}}$. Sketching a graphic of this described sampling distribution, we see the following.

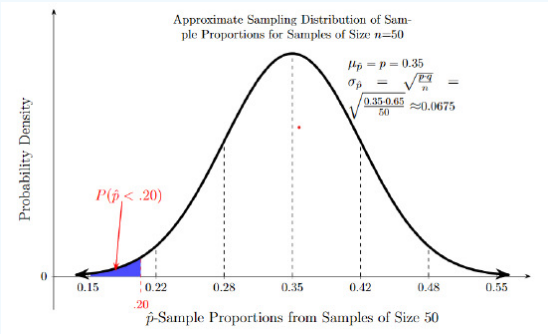


Figure 5.3.9 Approximate sampling distribution of sample proportions

Using our spreadsheet to compute the area/probability measure highlighted, we have:

$$P(\hat{p} < 0.20) = \text{NORM.DIST}(0.20, 0.35, \sqrt{\frac{0.35 \cdot 0.65}{50}}, 1) \approx 1.3083\%$$

We note that randomly selecting such a sample is not very likely, though also not impossible.

2. Determine the probability of randomly selecting eight hundred Kansas adults over 25 years old in which the sample's proportion will be within 2% of the actual population's proportion of 35%. That is, what proportion of samples of size eight hundred from this population of interest will produce a $\hat{p}$ measure between 33% and 37%?

Answer

We are again working in the same basic situation context, only with a larger sample size of 800.Again, turning to our developed theory we first note that $n \cdot p = 800 \cdot 0.35 = 280$ and $n \cdot q = 800 \cdot 0.65 = 520$ are both larger than 5. Therefore, we can reasonably approximate the sampling distribution of the possible $\hat{p}$ values as a normal distribution with a mean $\mu_{\hat{p}} = p = 0.35$ and a standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{0.35 \cdot 0.65}{800}}$. Sketching a graphic of this described sampling distribution, we see the following.

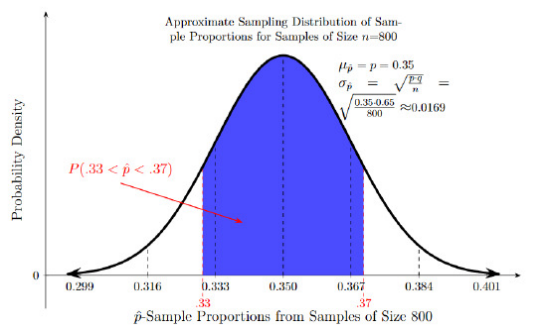


Figure 5.3.10 Approximate sampling distribution of sample proportions

We are interested in the proportion of samples in which the $\hat{p}$ values are within $2\% = 0.02$ of the population's measure of 35%. Thus, we need the area/probability measure in our distribution between $\hat{p}$ scale measures of $33\% = 0.33$ and $37\% = 0.37$. Using our technology to compute our area:

$$P(0.33 < \hat{p} < 0.37) \approx \text{NORM.DIST}(0.37, 0.35, \sqrt{\frac{0.35 \cdot 0.65}{800}}, 1) - \text{NORM.DIST}(0.33, 0.35, \sqrt{\frac{0.35 \cdot 0.65}{800}}, 1) \approx 0.882189 - 0.117811$$
$$\approx 76.4378\%$$

Around three-quarters of the samples of size 800 from the population of Kansas adults over 25 years old will produce sample proportion measures within 2 percentage points of the actual population proportion of 35%. Understanding such results gives us some confidence in using a single sample's measure from a sample of size 800 as a close approximation to what is happening in the population. We realize some samples will not meet this condition, but most will.

3. Regarding this situation involving Kansas adults over age 25, determine the probability of randomly selecting twelve Kansas adults over 25 years old in which at most 25% of the sample group have a bachelor's degree or higher. That is, determine $P(\hat{p} \leq 0.25)$.

Answer

We are again working in the same basic context as question 1. and 2. above, but we should also notice we are working with a somewhat small sample size. So checking our theory requirements, we first note that $n = 12$, $p = 35\% = 0.35$, and $q = 65\% = 0.65$, and thus our important requirement measures are $n \cdot p = 12 \cdot 0.35 = 4.2$ and $n \cdot q = 12 \cdot 0.65 = 7.8$. Since both are not larger than 5, we should NOT use the normal distribution for approximating the binomial probability distribution; we must go back to our discrete values table approach of the binomial probability distribution discussed in Section 4.3 of this text, with the needed adjustment to sample proportion as the random variable.

Building this table as per Section 4.3 concepts (for efficiency we use the BINOM.DIST function in a spreadsheet to produce the probability measures), and then converting to proportion measures on number of success as discussed above in this section, we produce the following distribution table:

Table 5.3.3 Binomial to sample proportion probability distribution

Number of Successes in n trials: X	**Proportion of Success $\hat{p} = \frac{x}{n}$**	**Probability $P(x) = P(\hat{p})$**
0	$\frac{0}{12} = 0.00 = 0\%$	0.5688%
1	$\frac{1}{12} \approx 0.08333 = 8.333\%$	3.6753%
2	$\frac{2}{12} \approx 0.16667 = 16.667\%$	10.8846%
3	$\frac{3}{12} = 25.000\%$	19.5365%
4	$\frac{4}{12} \approx 33.333\%$	23.6692%
5	$\frac{5}{12} \approx 41.667\%$	20.3920%
6	$\frac{6}{12} = 50.000\%$	12.8103%
7	$\frac{7}{12} \approx 58.333\%$	5.9125%
8	$\frac{8}{12} \approx 66.667\%$	1.9898%
9	$\frac{9}{12} = 75.000\%$	0.4762%
10	$\frac{10}{12} \approx 83.333\%$	0.0769%
11	$\frac{11}{12} \approx 91.667\%$	0.00753%
12	$\frac{12}{12} = 100.000\%$	0.000338%

This table is a representation of the $\hat{p}$-sampling distribution. We see that the sample proportion measure of 25% for samples of size 12 occurs when the binomial random variable is $x = 3$. Thus, by adding all associated probability measures when $\hat{p} \leq 0.25$, we find

$$P(\hat{p} \leq 0.25) = P(x \leq 3) \approx 0.346653 \approx 34.67\%.$$

Thus, about $34.67\%$of all the various possible samples of size 12 from the population of Kansas adults over 25 years old will produce a sample proportion value (representing the proportion of those with a bachelor's degree or higher) of at most 25%.

To see why this "check of requirements" was so important, we notice that if we had instead incorrectly used a normal distribution in this situation, we would have computed

$$P(\hat{p} \leq 0.25) \approx \text{NORM.DIST}\left(0.25, 0.35, \sqrt{\frac{0.35 \cdot 0.65}{12}}, 1\right) \approx 0.2333836 \approx 23.338\%$$

which is a significantly poor approximation value to the actual probability measure of the sampling distribution for proportions computed above of 34.67%.

4. Regarding this situation involving Kansas adults over age 25, determine the interval of sample proportions $\hat{p}$, which captures the central 95% of all possible proportion values from samples of size eight hundred. That is, determine $\hat{p}$-values a and b such that $P(a < \hat{p} < b) = 0.95$.

Answer

We are again working in the same basic situation as question 2. above, so we again go to the normal distribution modeling the sampling distribution of sample proportions: a normal distribution with a mean $\mu_{\hat{p}} = p = 0.35$ and a standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{0.35 \cdot 0.65}{800}}$. However, this time we have a central area/probability region of 95% and are looking for the boundary values on the $\hat{p}$-axis that captures this amount of area. Sketching a graphic of this described sampling distribution, we see the following:

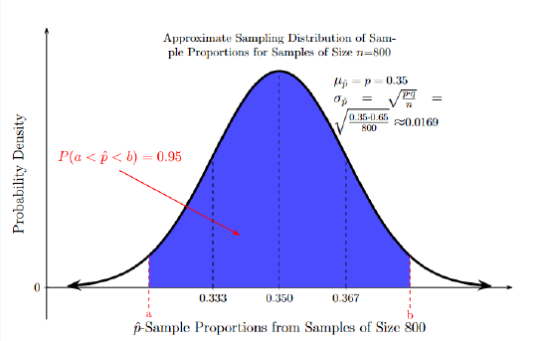


Figure $5.3.11$: Approximate sampling distribution of sample proportions

Reminding ourselves that we can find horizontal axis scale values in normal distributions tied to left area measures using our spreadsheet's NORM.INV function, we compute:

$$a = \text{NORM.INV}(0.025, 0.35, \sqrt{\frac{0.35 \cdot 0.65}{800}}) \approx 0.3169 = 31.69\%,$$

$$b = \text{NORM.INV}(0.975, 0.35, \sqrt{\frac{0.35 \cdot 0.65}{800}}) \approx 0.3831 = 38.31\%,$$

Thus, $P(0.3169 < \hat{p} < 0.3831) \approx 95\%$ or, stated in words, about 95% of all the various possible samples of size 800 from the population of Kansas adults over 25 years old will produce a sample proportion value (representing the proportion of those with a bachelor's degree or higher) between $31.69\%$and 38.31%.

We notice another implication from our work. Our boundary values of $31.69\%$and 38.31%each deviate from the population proportion value $p = 35\%$ by $|a - \mu_{\hat{p}}| = |b - \mu_{\hat{p}}| = |31.69\% - 35\%| = |38.31\% - 35\%| = 3.31\%$. This tells us that about 95% of random samples in this situation will produce a sample proportion measure $\hat{p}$ that is different from the population's proportion measure p by no more than $0.0331 = 3.31\%$. So most samples' proportions from samples of size 800 in this population are relatively close in value to the population's proportion, and only about 5% of samples will deviate from the population's proportion by more than 3.31%.

In summary, as long as certain requirements are met, we can often use normal distributions to analyze sampling distributions of sample proportions and understand how varied sample proportions can be within a specific binomial situation. As the Text Exercise $5.3.2.4$demonstrated, this will enable us to understand how we can infer, with some confidence, a population's proportion from a single random sample.

5.4: Sampling Distribution of Sample Variances - Optional Material

Learning Objectives

- State the expected value and variance of the sampling distribution of sample variances from a normally distributed parent population
- Discuss transforming the sampling distribution of sample variances to a χ^2-distribution
- Calculate probabilities of sample variances from normally distributed parent populations using χ^2-distributions

Review and Preview

At this stage, we have a relatively robust understanding of a sampling distribution, but we reiterate it once more within the context of sample variances. For a particular population, the sampling distribution of sample variances for a given sample size n is constructed by considering all possible samples of size n and computing the sample variances for each one. The values of the sample variances are the values that our random variable takes on. We then build the probability distribution with the understanding that the sampling method is simple random sampling. As such, we understand the sample variances as a random variable, which we typically treat as a continuous random variable.

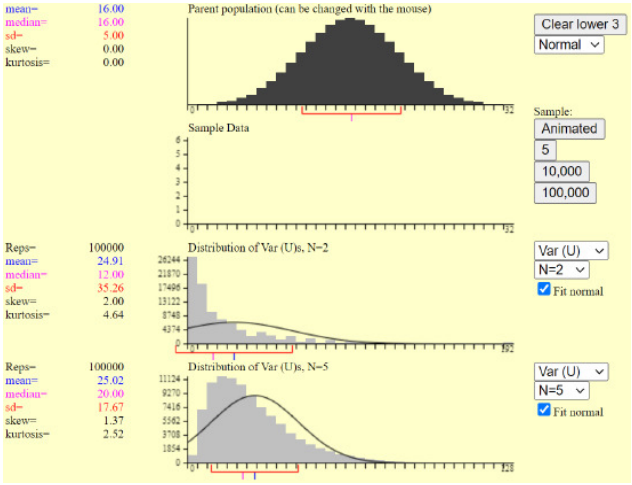
The construction of a sampling distribution is always the same. We have used this process for three sections. When considering sampling distributions of sample means, the Central Limit Theorem asserts that the sampling distribution becomes approximately normal as the sample size increases. This is true for any parent population. The smallest sample size at which the sampling distribution is approximately normal depends on the parent population.

We had something similar with the sampling distributions of sample proportions in the last section; if the sample was large enough to expect at least 10 observations with the given characteristic and at least 10 without the given characteristic, the sampling distribution was approximately normal. This was true for any parent population. Again, the smallest sample size at which the sampling distribution is approximately normal depends on the population proportion, but there is, theoretically, a sample size where it is approximately normal.

One may hypothesize that this is true for any sample statistic and population. Unfortunately, this is not the case when considering the sampling distribution of sample variances.

Sampling Distribution of Sample Variances

We have seen approximations to the sampling distribution of sample variances in the first section of this chapter. We are now using the same program to thoroughly explore the sampling distribution of sample variances. Using a normal parent population, we simulate the sampling distribution of sample variances through the same progression of sample sizes used in our previous development: $n = 2,\ 5,\ 10,\ 16,\ 20$, and 25. We have fit a normal curve to each distribution to emphasize that these sampling distributions are not approximately normal.



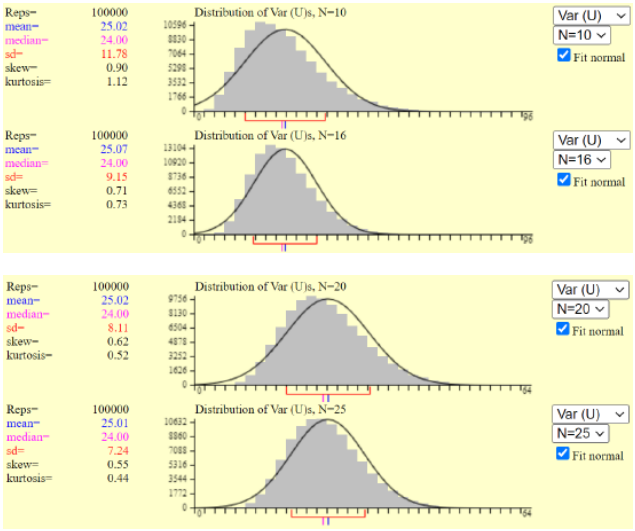


Figure 5.4.1: Sampling Distributions of Sample Variances for various sample sizes

Note that each sampling distribution of sample variances is centered about 25, the population variance. This happens because sample variance is an unbiased estimator of the population variance. Note that we have only estimated the sampling distribution of sample variances with a single example where the parent population is normal. We previously considered various parent populations. This reduction in scope is no accident; the method we describe works only for normally distributed parent populations. Some methods work for all parent populations, but those are beyond the scope of this course.

We provide, without proof, the expected value and standard deviation of the sampling distribution of sample variances in the case of a parent population that is normal with population standard deviation σ.

$$\mu_{s^2} = \mathrm{E}(s^2) = \sigma^2 \qquad \sigma^2_{s^2} = \mathrm{Var}(s^2) = \frac{2\sigma^4}{n-1}$$

Again, we emphasize that the distributions above are not represented well by normal curves. Recall that in chapter 4 we discussed a family of distributions that were positively skewed and followed a similar progression in shape as the degrees of freedom increased. We introduced the χ^2-distributions because they are at play in the sampling distributions of sample variances when the parent population is normally distributed. Consider the progression below (see that the figures above are frequency distributions with various scales, while the figures below are probability density functions all on the same scale).

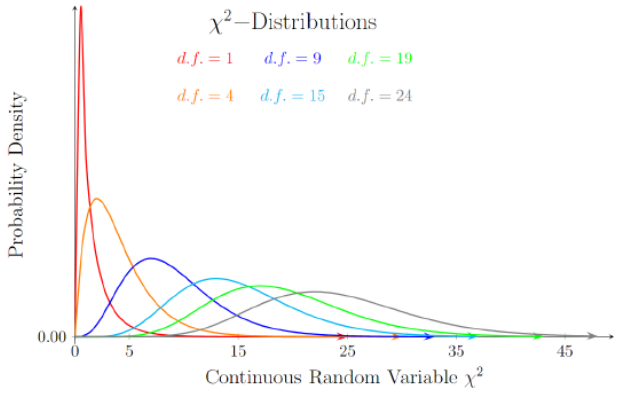


Figure 5.4.2: Chi-Square Distributions with Various Degrees of Freedoms

A rigorous development of the relationship between the sampling distribution of sample variances and the χ^2-distribution requires mathematical tools beyond the scope of this text. We provide only a brief exposition. Recall that the z-score can transform any normal distribution with mean μ and standard deviation σ into the standard normal distribution with mean 0 and standard deviation 1. This means we can study any normal distribution using the standard normal distribution. A similar situation is at play with the

sampling distribution of sample variances from a normal parent population. We must standardize the distribution and use technology to find the area. We must use a different transformation and probability distribution.

We will introduce the transformation within the familiar context of adult female heights. Recall that adult female heights are normally distributed with a mean of 64 inches and a standard deviation of 2.5 inches. We know that the population variance of adult female heights is $2.5^2 = 6.25$ square inches. We consider the sampling distribution of sample variances with a sample size of 10 and assess the probability of randomly selecting a sample of size 10 and getting a sample variance between 3 square inches and 9.5 square inches, $P(3 < s^2 < 9.5)$. Consider the following figures that illustrate the conversion.

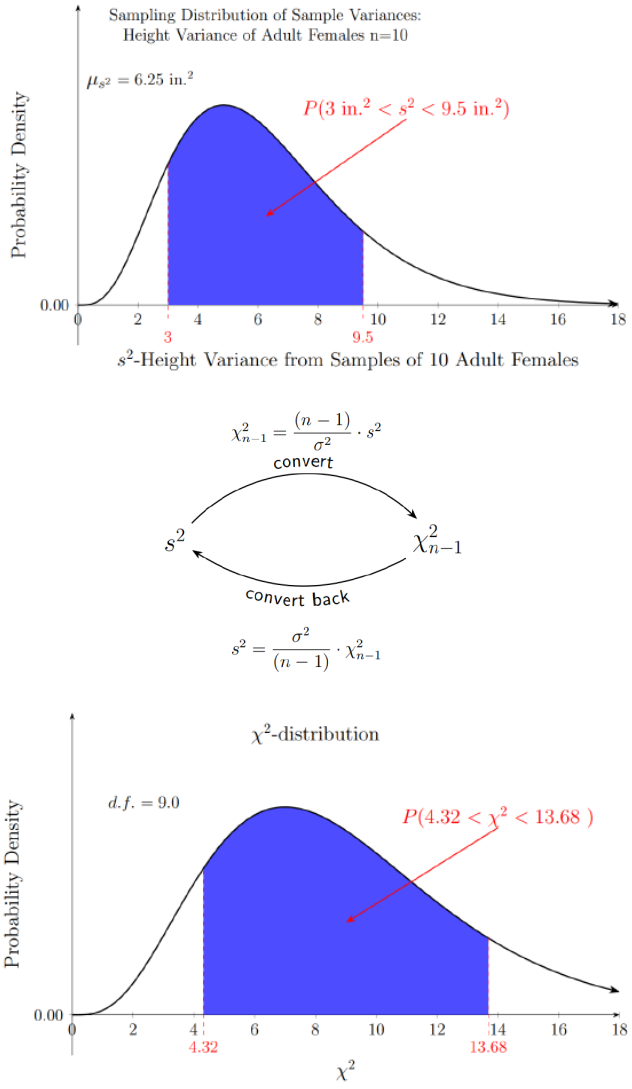


Figure 5.4.3: Transformation of a sampling distribution of sample variances to χ^2-distribution

Using the transformation $\chi^2_{n-1} = \frac{(n-1)}{\sigma^2} \cdot s^2$, with $n = 10$ and $\sigma^2 = 6.25$, we transform the sampling distribution of sample variances to the χ^2-distribution with 9 degrees of freedom, which we sometimes denote using $\chi^2_{10-1} = \chi^2_9$ to emphasize or provide the degrees of freedom. In this context, the degrees of freedom will always be $n-1$, one less than the sample size. We can compute $P(3 < s^2 < 9.5)$ by transforming the interval in terms of s^2 to an interval in terms of the $= \chi^2_9$ variable and computing the area using technology, as we did in chapter 4. Note that $\frac{(10-1)}{6.25} \cdot 3 = 4.32$ and $\frac{(10-1)}{6.25} \cdot 9.5 = 13.68$. So the interval, $3 < s^2 < 9.5$ gets transformed to $4.32 < \chi^2 < 13.68$. It is difficult to tell that these two areas are equal simply from the figure above, but it is indeed the case. We have plotted both distributions using the exact coordinates below. Compare the red and the blue areas to help convince yourself the claim of equal areas is possible/reasonable (for an interested reader, further exploration can be done using this Desmos comparison).

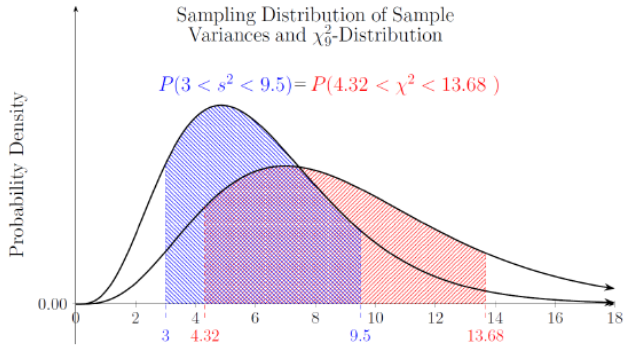


Figure 5.4.4: Sampling distribution of sample variances and χ^2-distribution plotted together to illistrate the preservation of area

We must introduce an accumulation function to calculate the area beneath χ^2-distributions. The function name and syntax may vary depending on the technology. We present a left-tailed accumulation function from Excel: CHISQ.DIST. The syntax in Excel is $=\text{CHISQ.DIST}(x \text{ value, degrees of freedom, cumulative})$. Since we are trying to find areas, we want cumulative to be marked true using TRUE or 1.

$$P(\chi^2_{n-1} < x) = \text{CHISQ.DIST}(x, n-1, 1)$$

With these tools, we may now compute $P(3 < s^2 < 9.5)$.

$$P(3 < s^2 < 9.5) = P(4.32 < \chi^2 < 13.68) = \text{CHISQ.DIST}(13.68, 9, 1) - \text{CHISQ.DIST}(4.32, 9, 1) \approx 75.4945\%$$

? Text Exercise 5.4.1

Adult IQ scores are thought to be normally distributed with a mean of 100 and a standard deviation of 15. Determine the probability that a random sample of 16 adults has a sample standard deviation less than 10.

Answer

We will not develop a sampling distribution of sample standard deviations since sample standard deviation is not an unbiased estimator of population standard deviation, even though sample variance is an unbiased estimator of population variance. If we are to proceed, we must translate our prompt into one that considers variance rather than standard deviation. If $s < 10$, then $s^2 < 100$ since standard deviation is non-negative. We are interested in computing $P(s^2 < 100)$ when the population is normally distributed, $n = 16$, and $\sigma^2 = 15^2 = 225$.

Let us transform our probability statement into one about a χ^2-distribution.

$$P(s^2 < 100) = P\left(\frac{n-1}{\sigma^2} \cdot s^2 < \frac{16-1}{225} \cdot 100\right) = P\left(\chi^2_{15} < \frac{20}{3}\right) = \text{CHISQ.DIST}(\frac{20}{3}, 15, 1) \approx 3.3752\%$$

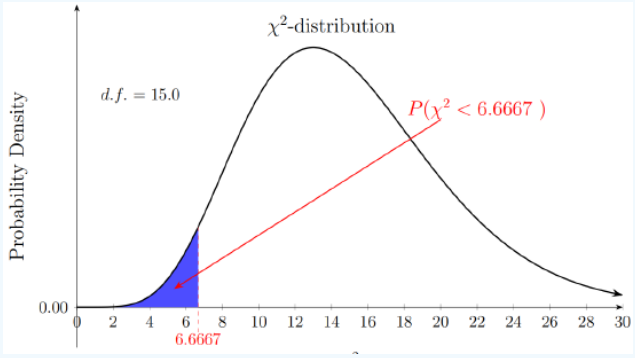


Figure 5.4.5: Sampling distribution of sample variances

CHAPTER OVERVIEW

6: Confidence Intervals

6.1: Introduction to Confidence Intervals

Learning Objectives

- Motivate the need for interval estimates
- Introduce and interpret confidence intervals
- Introduce margin of error
- Deduce information from confidence intervals knowing the general form

Review and Preview

Recall that in conducting inferential statistics, we are interested in understanding parameters, facts about a population, without having to do the work of studying the entire population. We want to study a sample and use the facts about it, the sample statistics, to estimate a population parameter. In the last chapter, we developed sampling distributions that connect the possible values of sample statistics with their probabilities. We found that, in general, we can approximate sampling distributions fairly well using continuous random variables. As such, we solidified our growing intuition that we should not expect a sample statistic computed from a simple random sample to be exactly equal to the population parameter. We expect there to be a difference between the two. The actual difference between a computed sample statistic and the population parameter must be unknown because the population parameter is unknown. Simply using a sample statistic as an estimate of the population parameter is insufficient. Instead, we estimate the population parameter by developing an interval estimate, called a **confidence interval**, based on the sample statistics and the sampling distribution.

In constructing an interval estimate, we hope to provide a meaningful range of values in which we feel confident that the population parameter is located. There are two competing desires here: the confidence that the population parameter is in the interval estimate, which we would like to be fairly high, and the length of the interval which we would like to be fairly small. But, as one might guess, increasing the confidence results in larger intervals. So, it is a balancing act, but luckily there is another factor at play that can help us manage both desires that we will study throughout this chapter. Let us begin our development of confidence intervals.

The Differences and Their Probabilities

Recall the text exercise about the grade distribution of statistics students for a particular instructor over several semesters. We computed the probability that the difference between the population mean 71% and the sample mean from a random sample of past students was less than 2%. For a sample size of 36, the probability that the sample mean fell within 2% of the population mean was 45.1494%.

We now make an important but seemingly trivial note. If the sample mean is within a certain distance of the population mean, then the population mean is within that same distance of the sample mean. We can translate the previous probability statement as such: the probability of randomly selecting a sample of size 36 so that the population mean is within 2% of the sample mean is about 45%. So, for about 45% of the samples of size 36, the population mean lies within 2% of the sample mean. Which is to say, for about 45% of the samples of size 36, the population mean lies in the interval $(\bar{x}-2, \bar{x}+2)$, where $\bar{x}$ is the computed sample mean from the randomly chosen sample. We might randomly select a sample of size 36 and compute a sample mean of 67%. In which case, the interval $(\bar{x}-2, \bar{x}+2) = (65, 69)$ does not contain 71 (see the red interval below). On the other hand, we might randomly select a sample of size 36 and compute a sample mean of 72.5, resulting in an interval of $(70.5, 74.5)$(see the dark blue interval below). We note that the population mean 71 does fall in this interval. Using a sample size of 36 with a difference of 2% results in the population parameter lying in the constructed interval about 45% of the time. We want to construct an interval that we are more confident in its ability to catch the parameter.

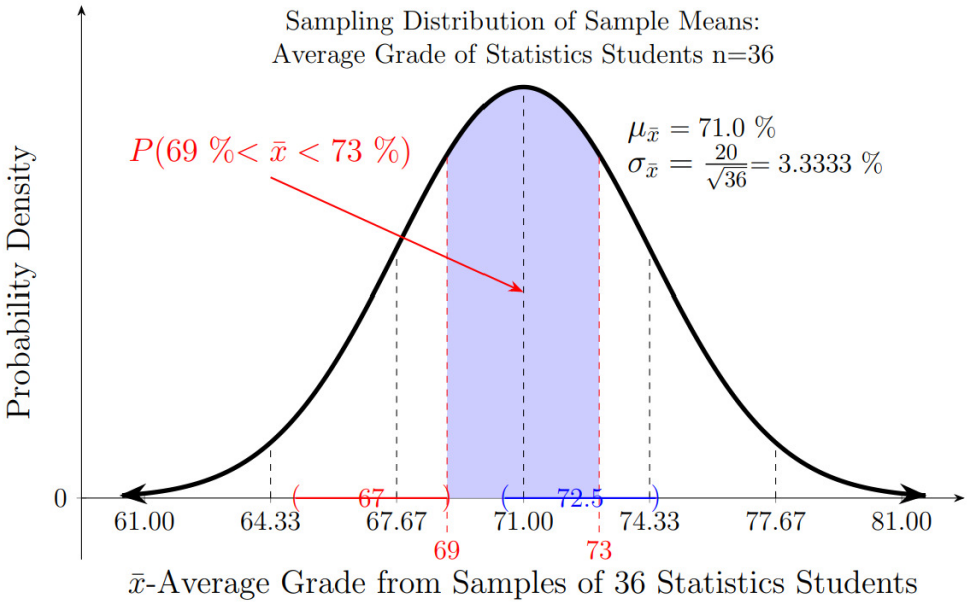


Figure 6.1.1: Sampling distribution of sample means $n = 36$

The text exercise progressed through various sample sizes to see the effect sample size had on the probability of the sample mean falling within 2% of the population mean. We noticed that the probability increased as the sample size increased. For a sample size of 288, the probability that the sample mean fell within 2% of the population mean was 91.0314%.Making a similar translation as before, the probability of randomly selecting a sample of size 288 so that the population mean is within 2% of the sample mean is about 91%. This means that for about 91% of the possible random samples of size 288 taken from our population, the interval constructed from our computed sample mean, $(\bar{x}-2, \bar{x}+2)$, will contain the population mean. If we did not know the population mean, we could not be sure which intervals successfully caught the population mean, but knowing that $91.0314\%$of the possible random samples of size 288 produce an interval containing the population mean elicits a certain confidence that most of the time we are successful in catching the population mean.

Herein lies an understanding of the name **confidence interval.** The **confidence level (CL)** of a confidence interval is the percentage of times (if we conducted random sampling repeatedly) that we would expect the population parameter to fall in our constructed interval. In general, the confidence level is set first and then the confidence interval is constructed to ensure the level of confidence. We now begin to fix the discrepancy in ordering.

Suppose, we want to construct a confidence interval at the level of 95% using a sample of size 288. Being within 2% of the population mean was insufficient to produce such a level of confidence. To increase our confidence, we must increase the distance the sample means are within. Will 3% be enough or 4%? Can we find the precise distance? Consider these questions in conjunction with the following figure.

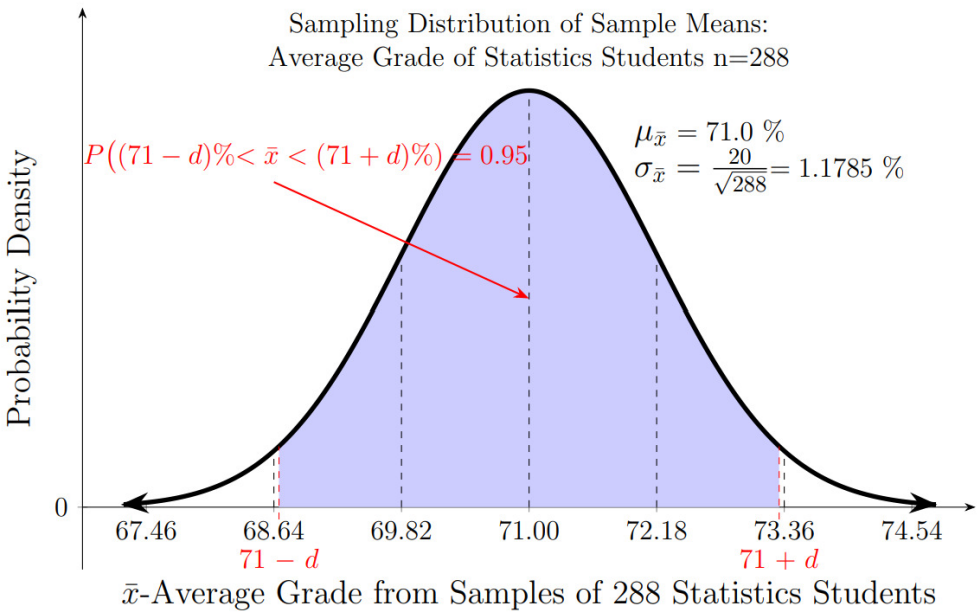


Figure 6.1.2: Sampling distribution of sample means $n = 288$

We are trying to find the distance d such that 95% of all samples of size 288 produce sample means that are within $d\%$ of the population mean. The shaded region has an area of 0.95, and the lower and upper bounds must be $d\%$ below and above the population mean, respectively. Using the fact that the total area under a probability density function is 1. We know that the total area of the white regions, the two tails, is $1-0.95=0.05$.. Notice that the two tails each have the same area since our probability density curve and our shaded region are symmetric about the mean. From this, we can deduce that the left tail, the white region on the left, has an area of $\frac{0.05}{2} = 0.025$. Now we know that $\text{NORM.INV}(0.025, 71, \frac{20}{\sqrt{288}}) \approx 68.6902$ is the lower bound of our shaded region, which also happens to be $71-d$. We have $d = 71 - \text{NORM.INV}(0.025, 71, \frac{20}{\sqrt{288}}) \approx 71 - 68.6902 = 2.3098$. Both 3% and 4% produce an interval wider than is necessary to attain 95% confidence.

95% of all samples of size 288 from our population produce sample means that are within 2.3098% of the population mean. If we construct an interval $(\bar{x} - 2.3098, \bar{x} + 2.3098)$ using the computed sample mean from a random sample of size 288, we would expect to catch the population mean in the constructed interval, 95% of the time. We would call 2.3098% our margin of error.

Confidence Intervals

Since the confidence level is a major driving force in constructing the confidence interval, the confidence level is given in conjunction with the confidence interval; a 95% confidence interval is a confidence interval constructed at the 95% confidence level. Recall that this indicates that the method of constructing the confidence interval, 95% of the time, produces an interval containing the population parameter. The remaining 5% of the time, the method fails to catch the population parameter; this rate of expected failure is often referred to as the **α value** (lowercase Greek letter alpha) of the confidence interval. The confidence level CL and α values are related to each other. When we construct a confidence interval, we either successfully catch the parameter or fail to catch the parameter. There are no other options. As such, the success rate plus the fail rate must be 1. Hence, $\text{CL} + \alpha = 1$. Common confidence levels are 90%, 95%, and 99%, but confidence levels can theoretically be any positive value less than 1 (100%). We can determine α for these common confidence levels since CL and α are complementary. So, we have α values of $0.1, 0.05,$ and 0.01, respectively.

If our confidence level is CL, our task is to answer the question: CL of the sample statistics are within what distance of the population parameter? This distance is called the **margin of error (ME)**, and it depends on the sampling distribution. An attentive reader will recognize a possible issue here. We knew the population mean was 71 when we computed d, the margin of error. How are we to compute such a distance if we do not know what the population parameter is? The solution is quite simple. We focused our study of sampling distributions primarily on two statistics: sample mean and sample proportion. Both of these statistics are unbiased estimators of their associated population parameters. This is important because we know that each respective sampling distribution's mean (expected value) is the associated population parameter. When the sampling distribution is approximately

normal, computing the margin of error reduces to computing the number of standard deviations from the mean that are necessary to gain the given confidence. We will discuss the specifics of computing the margin of error in later sections. For now, we recognize that the margin of error depends on the confidence level and the standard deviation of the sampling distribution (commonly referred to as the standard error). This method of confidence interval construction results in confidence intervals of the form: $(\text{sample statistic} - \text{margin of error}, \text{sample statistic} + \text{margin of error})$.

? Text Exercise 6.1.1

Use this information to answer the following questions. The state senate needs a two-thirds majority vote to override a governor's veto. A large random sample of senators was taken to estimate the percentage of senators who support overriding the veto. The 90% confidence interval for proportions $(0.67, 0.73)$ was constructed using the sample data and the method discussed above.

1. Explain the meaning of the 90% confidence level and the resulting confidence interval in the context of the problem.

Answer

The 90% confidence level indicates that the method of constructing the confidence interval catches the population proportion 90% of the time. So, we are 90% confident that the population proportion p falls somewhere between 67% and 73%.

2. Based on the confidence interval, would you recommend that a proponent of the override motion initiate a vote?

Answer

Since the lower boundary of the confidence interval is 67%, which is greater than two-thirds, the proponent of the override can feel confident that the Senate can override the governor's veto, although it appears that the margin of victory will not be very large. We, therefore, recommend that the proponent initiate the vote.

3. Determine the computed sample proportion $\hat{p}$ and margin of error (ME).

Answer

Because this line of reasoning applies to any confidence interval constructed using the method discussed above, we answer this question first in generality and then in particular. Since the confidence interval was constructed using the method discussed above, we know that the lower bound is equal to the sample statistic minus the margin of error and that the upper bound is equal to the sample statistic plus the margin of error. So we have the following system of equations.

$$\begin{aligned} \text{lower bound} &= \text{sample statistic} - \text{margin of error} \\ \text{upper bound} &= \text{sample statistic} + \text{margin of error} \end{aligned}$$

We can eliminate the margin of error by adding the two equations together.

$$\text{lower bound} + \text{upper bound} = 2 \cdot \text{sample statistic}$$

We can eliminate the sample statistic by taking the difference of the two equations.

$$\text{upper bound} - \text{lower bound} = 2 \cdot \text{margin of error}$$

We have developed the following formulas.

$$\text{sample statistic} = \frac{\text{lower bound} + \text{upper bound}}{2} \qquad \text{margin of error} = \frac{\text{upper bound} - \text{lower bound}}{2}$$

This confirms that the sample statistic is the midpoint of the confidence interval and that the margin of error is half of the length of the confidence interval. So $\hat{p}$ is the midpoint between 0.67 and 0.73 which is 0.70, and ME is the distance from the midpoint to an end or half of the interval length which is 0.03.

Note: Reading Symbols with Meaning

Reading mathematical symbols with meaning is an important skill to develop in a quantitatively and symbolically driven society. As we have seen in the development of confidence intervals, the margin of error, ME, is the distance such that the confidence level, CL, produces sample statistics within that distance of the population parameter. We can understand this equivalently as the maximum distance a sample statistic could be from the population parameter and still be captured in the confidence interval. Since we are conducting simple random sampling in the context of this course, we can understand that statement as the probability of randomly selecting a sample of a given size that produces a sample statistic within ME of the population parameter is CL. This can be expressed with symbols quite elegantly. The symbols for statistics and parameters differ based on the context; so, for now, let us remain in the context of means. The statements made above can be expressed as $P\left(|\bar{x}-\mu|<\mathrm{ME}\right)=\mathrm{CL}$.

Read $|\bar{x}-\mu|$, the absolute value of the difference between the sample mean and population mean, as the distance between the sample mean and population mean.

Read $|\bar{x}-\mu|<\mathrm{ME}$ as the distance between the sample mean and population mean is less than the margin of error.

Read $P\left(|\bar{x}-\mu|<\mathrm{ME}\right)$ as the probability that the distance between the sample mean and population mean is less than the margin of error.

Read $P\left(|\bar{x}-\mu|<\mathrm{ME}\right)=\mathrm{CL}$ as the probability that the distance between the sample mean and population mean is less than the margin of error is the confidence level.

It is important to remember that the random variable at play is the sample statistic because the underlying random experiment is conducting a random sample of a given size.

Text Exercise 6.1.2

Read the following mathematical expression with meaning: $P\left(|\hat{p}-p|>\mathrm{ME}\right)=1-\mathrm{CL}=\alpha$

Answer

The probability that the distance between the sample proportion and population proportion is greater than the margin of error is one minus the confidence level, which is the alpha value.

6.2: Confidence Intervals for Proportions

Learning Objectives

- Recognize that a proportion of a random sample proportion is an estimation of the related proportion measure of the population
- Develop and apply the margin of error measure for using the proportion of a sample proportion to estimate the proportion of the population
- Develop and apply the confidence interval measures for using the proportion of a sample to estimate the proportion of the population
- Develop and apply sample size measures to control margin of error

Review and Preview

As stated numerous times before, an important area of inferential statistics is the ability to use a single measure from a sample to predict the related measure for the entire population (such as using the mean of a sample to predict the mean of the population or using the proportion from a sample to predict the proportion of the population.) In the previous Section 6.1, we discussed the general concepts of margin of error, confidence intervals, confidence levels, and α value; all of which are important measures of inferential statistics. We now focus on the specific situation of using a proportion measure from a random sample to predict a proportion measure from the population.

To further review, we remind ourselves of Section 5.3 and the sampling distribution of sample proportions where we noted that different samples of a specific chosen size, n, produce a collection of various sample proportion measures $\hat{p}$. In our past investigations, most if not all of the various sample proportion measures were not the same value as the population's proportion, that is $\hat{p} \neq p$. It was also important for us to recognize that in the large collection of various $\hat{p}$ values, that under certain restrictions, the distribution of $\hat{p}$ values formed an approximately normal distribution. (The restrictions required that $n \cdot p > 5$ and $n \cdot q > 5$, both of which tend to be easily met if working with large sample sizes.) Furthermore, this sampling distribution's mean value will be the same as the population's proportion value and the spread (standard deviation) in the sampling distribution is smaller than the standard deviation of the population. In notational form, we designated this with $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}}$.

As one final review note, we re-examine the third part of Text Exercise \(5.3.\2.) In that exercise, we found the central interval in the sampling distribution that contained 95% of possible sample proportion results. That is, we found within the given context how far away (the margin of error) from the population proportion's value 95% of the various samples' proportions would be.

Now we use these previous findings to develop a routine method for building a confidence interval in the proportion measure situation.

Sampling Distribution of Sample Proportions and Confidence Intervals

Let us begin in a specific context to help frame our work. Suppose that we are interested in predicting the proportion of the U.S. adult population which has received the latest flu vaccine. Naturally, we would not be able to ask every U.S. adult due to the population size and likely limited resources/finances to collect such data. However, it would be reasonable for us to randomly contact $1,000$ such adults in the United States and determine which of those had and which had not taken the latest vaccine. Suppose 735 of those had received the vaccine; then this one collected sample had a proportion measure of $\hat{p} = \frac{735}{1000} = 0.735 = 73.5\%$. Naturally, we do not claim the population's proportion, p, is the same value, but our work with sampling distributions should convince us that we can expect the population's measure to be reasonably close to this sample's measure. This predictable sampling distribution of sample proportions allows us to consider a random sample's proportion, $\hat{p}$, to be a valid ***point estimate*** of the population's proportion; after all, the sampling distribution shows that most of the time a random sample's proportion will be "close" to the population's proportion measure. However, we need to have a measure for "close".

Due to the predictable sampling distribution of sample proportions (shown in Figure $6.2.1$ below), we will determine a measure of "close" by choosing a confidence level (CL) value, such as 95%. Our measure of "close" will be a calculated margin of error measure designated as ME. Recall that the distribution below shows that most random samples (in fact the percentage given by our choice of CL) will produce $\hat{p}$ values that fall within the ME distance of the actual population's proportion p which is at the center of the sampling distribution. As long as we choose a large CL value, we have a very good chance that our one collected random

sample's proportion will fall on the horizontal axis scale under the blue region. (Yes, it is possible that our random sample's $\hat{p}$ will not be within this group, but the probability of such an outcome is only $\alpha = 100\% - \text{CL}$--the α value discussed in Section 6.1. Once more, we can see that if we keep our choice of CL close to 100%, then α will be small: close to 0%.

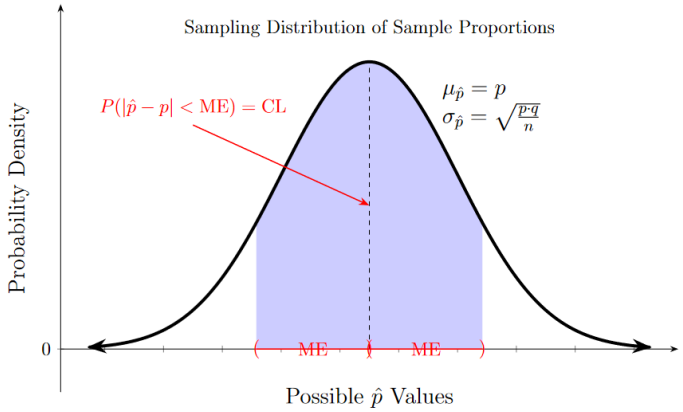
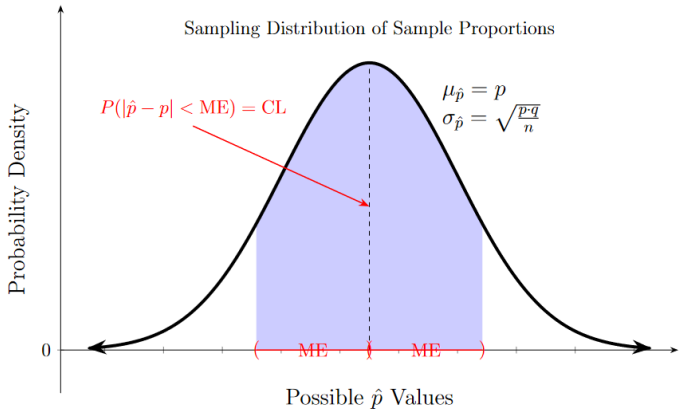


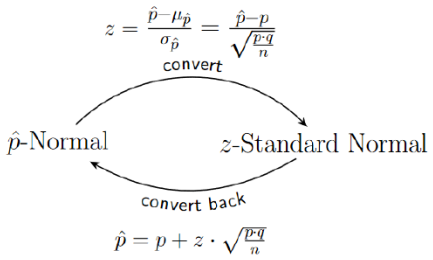
Figure 6.2.1: Sampling distribution of sample proportions

Now in relation to our given situation of estimating the proportion of all U.S. adults that have received the flu vaccine by using a sample of size $n = 1000$, we note that we do not know p and so, unlike our previous work in Chapter 5, we cannot determine the scaling of our horizontal axis in the natural scale of proportion measures. We address this in the next subsection on determination of the margin of error ME value.

Determining the Margin of Error in the Proportion Situation

We now use our powerful standardization feature on normal distributions from Section 4.5; where any normal distribution can be converted in scale to the standard normal distribution. Consider the following figures illustrating the transformation process to our normal distribution of sample proportions.





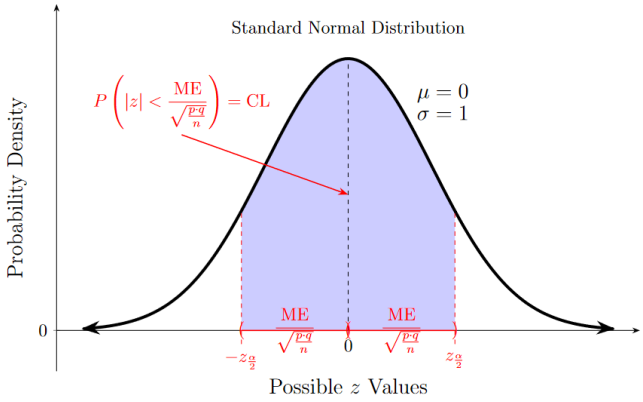


Figure 6.2.2: Sampling distribution of sample proportions transformed into the standard normal distribution

Notice that under the standard normal transformation process, the margin of error is scaled by a factor of one over the standard deviation of the sampling distribution; this occurs since we divided by the standard deviation value of $\sqrt{\frac{p \cdot q}{n}}$ in our scale transformation. Also, since we are now considering the standard normal distribution, we know the mean and standard deviation of the distribution. Therefore, using our computation technology, we can find the boundary values in the z-scale that will produce the desired confidence. These are represented by $\pm z_{\frac{\alpha}{2}}$ in the figure above. We call these points **critical z-values** or simply **critical values** in this process; these are completely determined once we have a chosen confidence level. Note that $z_{\frac{\alpha}{2}}$ represents the z-scale value where the area under the standard normal distribution to the right is $\frac{\alpha}{2}$ and $-z_{\frac{\alpha}{2}}$ represents the z-scale value where the area under the standard normal distribution to the left is $\frac{\alpha}{2}$. In converting back to the proportions' sampling distribution with $\hat{p} = p + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$, or in related equivalent form of $\hat{p} - p = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$, we should recognize something important. These critical z-values are telling us how many standard deviations of the sampling distribution we must differ from the mean of that distribution to capture the chosen CL percentage of sample results. That is, we have a measure of our "closeness" by $\text{ME} = \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot q}{n}}$.

We do have an issue in this computation of the margin of error since it needs the value of p, yet the value of p is unknown and what we are trying to estimate. However, since samples' proportions tend to be close in value to the population proportion, we will use our sample's proportion measure in the calculation. That is, we will find the margin of error measure of closeness by $\text{ME} \approx \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$. For practical purposes, the use of $\hat{p}$ and $\hat{q}$ instead of p and q is reasonable as long as we meet the large-sample requirements of $n \cdot \hat{p} > 5$ and $n \cdot \hat{q} > 5$. It is worth noting that using $\hat{p} \cdot \hat{q} \approx p \cdot q$ is not the same as using $\hat{p} \approx p$; there is smaller error in the former than the latter. One illustration of this, which is developed further later in the section, is the fact that $p \cdot q$ is never larger than 0.25. For example, if $p = 0.4$ and we get $\hat{p} = 0.3$ then $p \cdot q = 0.24$ and $\hat{p} \cdot \hat{q} = 0.21$; even large differences between p and $\hat{p}$ may still yield small differences between $p \cdot q$ and $\hat{p} \cdot \hat{q}$. This is worth observing so that it does not seem as if we are chasing our own tail. Our goal is to estimate p; it would be pointless to use a formula to do so if the formula implicitly used $p \approx \hat{p}$. If we are concerned with the accuracy of our error measure, we can be more conservative and instead require $n \cdot \hat{p} > 10$ and $n \cdot \hat{q} > 10$ as discussed in Section 5.3. By using larger sample sizes, we can be more assured of our theory and hence in the validity of our measures produced by this theory. One should not use the above ideas on proportions measures if working with small sample sizes. We summarize the above in the following.

Margin of Error in Sample Proportions

Given a desire to estimate a population proportion measure p using a simple random sample's proportion $\hat{p}$ in which the following conditions are known or reasonably believed to exist:

- the requirements for a binomial distribution are met with a sample size of n
- the requirements of $n \cdot \hat{p} > 5$ and $n \cdot \hat{q} > 5$ are met
- a confidence level of CL has been chosen and hence $\alpha = 100\% - \text{CL}$

then the margin of error in using the random sample's proportion measure is measured by

$$\text{ME} = \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

where $\pm z_{\frac{\alpha}{2}}$ are the two critical values capturing the center $\text{CL}\%$ of the standard normal distribution.

If being more conservative in our approach, we may instead use requirements of $n \cdot \hat{p} > 10$ and $n \cdot \hat{q} > 10$.

With this theory in place, we now apply this to our specific context of the flu vaccine. We recall that we were interested in estimating the proportion of all U.S. adults who had taken the most recent flu vaccine. We had collected a random sample of $1,000$ adults in which 735 had taken the vaccine, producing $\hat{p} = 0.735 = 73.5\%$. We note that the requirement for a binomial distribution are met with this context in relation to samples of size $n = 1,000$ and that $n \cdot \hat{p} = 1000 \cdot 73.5\% = 735 > 5$ and $n \cdot \hat{q} = 1000 \cdot 26.5\% = 265 > 5$.

Next, we do expect the actual population's proportion to be close to this 73.5% value due to our sampling distribution theory, but we need a measure of how close: a measure of the likely margin of error in the sample's result. To do so, we first must set a confidence level, say we choose $\text{CL} = 95\%$. This means that this process will produce an interval which contains p 95% of the time. Then, to determine this margin of error, we proceed to the standard normal distribution to find the associated critical z-values tied to a central area of 95% and left/right tail areas of $\frac{\alpha}{2} = 2.5\%$, illustrated below in Figure $6.2.3$.

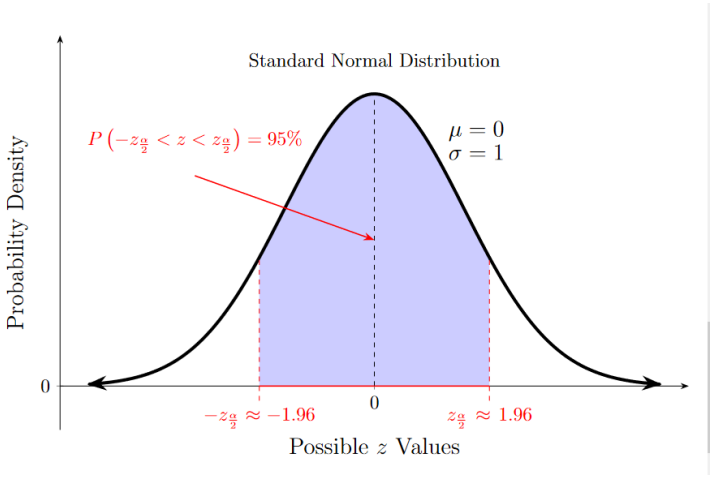


Figure $6.2.3$: Standard normal distribution for a 95% confidence level

Using our approach of Section 4.6, we find these critical z-scores using our spreadsheet's NORM.S.INV function:

$$\begin{aligned} \text{left critical } z_{\frac{\alpha}{2}} &= \text{NORM.S.INV}(0.025) \approx -1.95996 \\ \text{right critical } z_{\frac{\alpha}{2}} &= \text{NORM.S.INV}(0.975) \approx 1.95996 \end{aligned}$$

Of course, as seen in earlier work, since the standard normal distribution is symmetric about its mean scale value of 0, we need not actually compute both critical $z-$values as both will be the same sized value, just one negative and the other positive.

This now lets us determine our margin of error as tied to this chosen confidence level of 95% :

$$\begin{aligned} \text{ME} &= \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \\ &\approx \pm 1.95996 \cdot \sqrt{\frac{0.735 \cdot 0.265}{1000}} \\ &\approx \pm 1.95996 \cdot 0.013956 \approx \pm 0.02735 = \pm 2.735\% \end{aligned}$$

Thus we have 95% confidence that our one random sample's proportion of $\hat{p} = 73.5\%$ is no more than 2.735% away from the population's actual proportion measure p. That is, in assuming our one collected sample's proportion is one of the central 95% of possible sample proportion values that can occur from samples of size $1,000$, then our sample's proportion will be found on the horizontal scale somewhere below the shaded region, no more than 2.735% from the actual true population's proportion, as illustrated in Figure $6.2.4$ below.

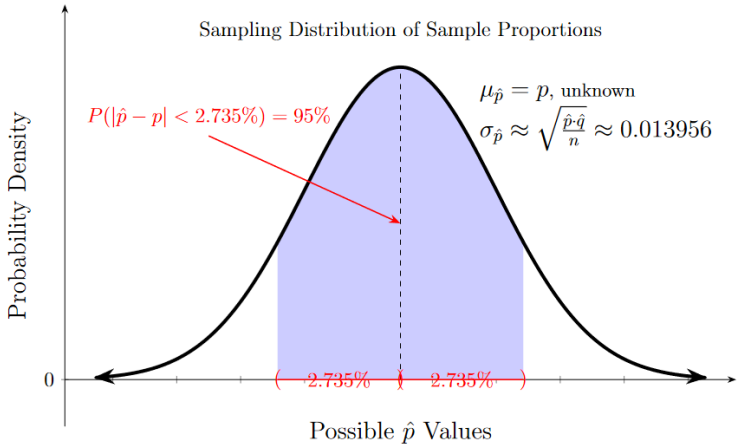


Figure 6.2.4: Illustration of margin of error in a sampling distribution

In the above computation of the margin of error measure, one should note that, although the confidence level was $\text{CL} = 95\%$ and complement alpha level was thus $\alpha = 5\%$, in determination of the critical z-values within the NORM.S.INV function, neither of these two numbers were directly used. Instead, since the spreadsheet's function requires use of only a left-area measure, we instead had to use $\frac{\alpha}{2} = 2.5\%$ and its complement measure of $1 - \frac{\alpha}{2} = 97.5\%$ within the spreadsheet function. This is a technology computational requirement that must be recalled when constructing these measures.

As a final summary of our specific example, we are able to state that we have 95% confidence that the true proportion of the U.S. adult population that had taken the most recent flu vaccine is approximately 73.5% with no more error than 2.735%. This now easily leads us to the final concept of this section, the **confidence interval** for the proportion situation.

Constructing Confidence Intervals for Proportions

Once we have a margin of error measure determined, we easily construct the confidence interval for the population proportion.

$$(\hat{p} - \text{ME}, \hat{p} + \text{ME}) = \left(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}\right)$$

Or, instead of using algebraic interval notation, we may instead indicate the confidence interval as follows.

$$\hat{p} \pm \text{ME} = \hat{p} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

So in our flu vaccine context, we have a confidence interval of $73.5\% \pm 2.735\%$ or equivalently $(0.735 - 0.02735, 0.735 + 0.02735) = (0.707646, 0.762354) = (70.7646\%, 76.2354\%)$.

This allows us to state that we are 95% confident that the actual proportion of the U.S. adult population that had taken the most recent flu vaccine is between $70.7646\%$and 76.2354%,or equivalently $73.5\% \pm 2.735\%$.

Let us try a few more text exercises using the same theory but in varied contexts.

? Text Exercise 6.2.1

Use our theory on margin of error and confidence intervals established above, determine the following.

1. A state's department of education is interested in the proportion of all eighth-grade students in their state that will score at less-than-proficient in math on a national assessment. A random sample of 450 eighth-grade students from the state were given the national assessment and 345 of those students scored less-than-proficient in math. Develop an appropriate estimate from this information for the department of education, including the margin of error and related confidence interval based upon a choice of a 90% confidence level. Include a final concluding statement with the developed confidence interval.

Answer

We proceed by first developing the sample's proportion measure:

$$\hat{p} = \frac{345}{450} \approx 0.76666667 \approx 76.67\%$$

Thus, in the sample, about $\hat{p} = 76.67\%$ of the sampled eighth-grade students scored less-than-proficient on the national assessment and $\hat{q} \approx 23.33\%$ scored above less-than-proficient. We also note that we meet the basic requirements for our theory since $n \cdot \hat{p} = 345$ and $n \cdot \hat{q} = 105$ are both well above 5 and the situation is based on a random sample and a binomial experiment.

We now use this sample measure as a "best estimate" for the population's proportion, but also need to determine the possible likely margin of error in this estimate. We continue by developing $\text{ME} = \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$. Since working with a chosen confidence level of 90%, then $\alpha = 10\%$, leading to determination of the following critical z-values.

$$\pm z_{0.05} = \pm \text{NORM.S.INV}(0.05) \approx \pm 1.64485$$

The statistical margin of error is

$$\begin{aligned} \text{ME} &= \pm z_{0.05} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \\ &\approx \pm 1.64485 \cdot \sqrt{\frac{0.7667 \cdot 0.2333}{450}} \\ &\approx \pm 1.64485 \cdot 0.019938 \approx \pm 0.0328 = \pm 3.28\% \end{aligned}$$

Thus, based upon 90% confidence, we have at most a 3.28% margin of error in this sample estimate; leading to a confidence interval of $76.67\% \pm 3.28\%$, or in interval notation, $(0.7339, 0.7995) = (73.39\%, 79.95\%)$.

As a final summary, we are able to state that we have 90% confidence that the true proportion all eighth-grade students in this state that will score at less-than-proficient in math on the state assessment is approximately 76.67% with no more error than 3.28%; we are 90% confident that the true population proportion p falls somewhere between 73.39% and 79.95%.

2. A marketing researcher is interested in the proportion of European consumers who are aware of a U.S. branded product. A random sample of 375 European consumers were asked if they recognized the U.S. branded product; 75 stated they knew of the product. Develop an appropriate estimate of the proportion of all European consumers who are aware of the U.S. branded product from this information for the researcher, including the margin of error and related confidence interval based upon a choice of a 99% confidence level.

Answer

We again first develop the sample's proportion measure:

$$\hat{p} = \frac{75}{375} = 0.2000 = 20\%$$

Thus, in the sample, $\hat{p} = 20\%$ of the sampled Europeans were aware of the U.S. branded produce and $\hat{q} = 80\%$ were not aware. We also note that we meet the basic requirements for our theory since $n \cdot \hat{p} = 75$ and $n \cdot \hat{q} = 300$ are both well above 5 and the situation is based on a random sample and a binomial experiment.

We use this sample proportion as a "best estimate" for the population's proportion, but must determine the possible likely margin of error in this estimate. Since we are working with a chosen confidence level of 99%, then we have $\alpha = 1\%$, leading to determination of the following critical z-values of

$$\pm z_{0.005} = \pm \text{NORM.S.INV}(0.005) \approx \pm 2.57583$$

Thus, the statistical margin of error is

$$\begin{aligned} \text{ME} &= \pm z_{0.005} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \\ &\approx \pm 2.57583 \cdot \sqrt{\frac{0.20 \cdot 0.80}{375}} \\ &\approx \pm 2.57583 \cdot 0.020656 \approx \pm 0.0532 = \pm 5.32\% \end{aligned}$$

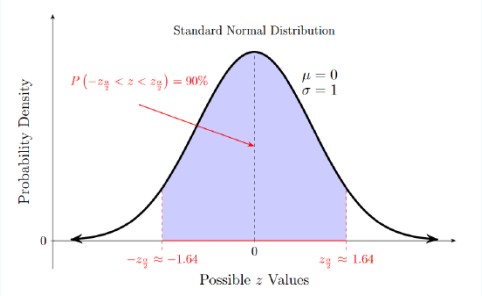
Based upon 99% confidence, we have at most a 5.32% margin of error in this sample estimate; leading to a confidence interval of $20\% \pm 5.32\%$, or in interval notation, $(0.1468, 0.2532) = (14.68\%, 25.32\%)$.

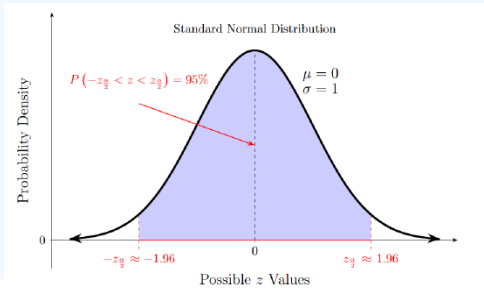
As a final summary, we are able to state that we have 99% confidence that the true proportion all Europeans that recognize the U.S. branded product is approximately 20% with no more error than 5.32%; we are 99% confident that the actual population proportion p falls somewhere between 14.68% and 25.32%.

3. What will happen to the margin of error if one increases the desired level of confidence?

Answer

Since the confidence level CL is tied to us assuming our random sample's proportion is within the central CL% of the sampling distribution or its standardization, we need only recognize what happens to our horizontal axis interval in relation to any adjustment of the level. Using a confidence level of 90 first and then of 95%, we can visually reason that increasing the confidence level increases the size of the horizontal axis interval (and hence the size of the related critical z-values) as is illustrated in the diagrams below.





Such an increase in the chosen confidence level will then cause the margin of error measure to be larger and hence the confidence interval to be wider. So, choosing to increase only the desired level of confidence (while also not changing any other option) will cause a larger margin of error. The ethical researcher will always set the confidence level before beginning the statistical analysis (not set after some statistical work just to force a smaller margin of error.) Those aware of this will also notice when research sets an unusually low confidence level, possibly in an attempt to narrow the margin of error in sample results so as to mislead consumers of the research.

4. What will happen to the margin of error if one decreases the sample size used to produce the sample proportion estimate?

Answer

Since $\text{ME} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, we have that n is in the denominator. Using our number sense within simple arithmetic, we see that as n decreases, our denominator in our sampling distribution's standard deviation measure, $\sqrt{n}$, also decreases. When dividing by smaller and smaller numbers, the result of the quotient is larger and larger (for example, $\frac{1}{500} = 0.002$, $\frac{1}{50} = 0.02$, $\frac{1}{5} = 0.2$, and so on. As we decrease the denominator, the value of our fraction increases getting closer and closer to 1. Thus, the margin of error becomes larger as the sample size gets smaller. This should match our natural number sense that smaller samples are more likely to produce statistics which deviate more from the population parameter in comparison to larger samples.

Sample Size Determination in Confidence Intervals on Proportions

Based upon part 4 of the last text exercise group, we notice that sample size choice plays some role in controlling the magnitude of the margin of error. We can apply a bit of algebraic manipulation to develop a formula allowing us to pre-predict the sample size

needed to control the margin of error within any specific chosen level of confidence. Using our developed margin of error formula,

$$\text{ME} = \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}},$$

we note that we can solve this algebraic formula for n, as illustrated below.

$$\begin{aligned}
(\text{ME})^2 &= \left(\pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}\right)^2 && - \text{ square both sides} \\
\text{ME}^2 &= \left(z_{\frac{\alpha}{2}}\right)^2 \cdot \frac{\hat{p} \cdot \hat{q}}{n} && - \text{ simplify} \\
n \cdot \text{ME}^2 &= \left(z_{\frac{\alpha}{2}}\right)^2 \cdot \frac{\hat{p} \cdot \hat{q}}{\cancel{n}} \cdot \cancel{n} && - \text{ multiply both sides by } n \\
\frac{n \cdot \cancel{\text{ME}^2}}{\cancel{\text{ME}^2}} &= \left(z_{\frac{\alpha}{2}}\right)^2 \cdot \hat{p} \cdot \hat{q} \cdot \frac{1}{\text{ME}^2} && - \text{ divide both sides by } \text{ME}^2 \\
n &= \left(\frac{z_{\frac{\alpha}{2}}}{\text{ME}}\right)^2 \cdot \hat{p} \cdot \hat{q} && - \text{ simplify}
\end{aligned}$$

So we have developed a related formula that will tell us how large of sample we need once we have chosen a confidence level (so we can determine the critical z-value), a margin of error size, and some previous study's sample results (so we have values for $\hat{p}$ and $\hat{q}$.) It would be nice to eliminate the requirements of a previous study, and we can do so if we take just a brief time to notice that the product $\hat{p} \cdot \hat{q}$ is predictable. Recall that $\hat{q}$ is the complement of $\hat{p}$, so as illustrated by the table of values below.

Table 6.2.1: Products of proportion values and their complements

$\hat{p}$	$\hat{q}$	Product $\hat{p} \cdot \hat{q}$
0.00	1.00	$0.00 \cdot 1.00 = 0.00$
0.10	0.90	$0.10 \cdot 0.90 = 0.09$
0.20	0.80	$0.10 \cdot 0.90 = 0.16$
0.30	0.70	0.21
0.40	0.60	0.24
0.50	0.50	0.25
0.60	0.40	0.24
⋮	⋮	⋮
1.00	0.00	0.00

We can inductively reason that the maximum product is $0.50 \cdot 0.50 = 0.25$. So a required sample size in the proportion situation can be found without a preliminary study by our developed formula given by:

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot 0.50}{\text{ME}}\right)^2 = \left(\frac{z_{\frac{\alpha}{2}}}{\text{ME}}\right)^2 \cdot 0.25$$

The above leads to the following key findings.

Sample Size for Estimation of Population Proportions

Given a desire to estimate a population proportion measure p using a simple random sample's proportion $\hat{p}$ in which the required conditions are to be met, then the sample size needed to meet a confidence level of CL% and margin of error of no more than ME can be found by the following computations:

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{\text{ME}}\right)^2 \cdot \hat{p} \cdot \hat{q} \quad - \text{ if a preliminary value of } \hat{p} \text{ is known}$$

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{\text{ME}}\right)^2 \cdot 0.25 \quad - \text{ if no preliminary value of } \hat{p} \text{ is known}$$

Now we apply these sample size concepts within a few exercises.

Text Exercise 6.2.2

Using our sample size findings above, determine the following.

1. A state's department of education is interested in the proportion of all eighth-grade students in their state that will score at less-than-proficient in math on a national assessment. A researcher is interested in controlling the margin of error to no more than 1.5% while working under a 95% confidence level. A previous study from three years ago produced a sample proportion measure of $\hat{p} = 58\%$. What size sample is required for the researcher to meet the desired conditions?

Answer

We proceed by applying our developed sample size formula in which we need the margin of error to be at most $\text{ME} = 1.5\%$ and also happen to have a preliminary value of $\hat{p} = 0.58$ known. First, we must determine the critical z-scores $z_{\frac{\alpha}{2}}$ tied to the prescribed confidence level of 95%. Therefore $\alpha = 5\%$ and $\frac{\alpha}{2} = 2.5\%$.

$$z_{0.025} = \pm\text{NORM.S.INV}(0.025) \approx \pm 1.95996.$$

Hence, our sample size calculation is

$$\begin{aligned} n &= \left(\frac{z_{\frac{\alpha}{2}}}{\text{ME}}\right)^2 \cdot \hat{p} \cdot \hat{q} \\ &= \left(\frac{\pm 1.95996}{0.015}\right)^2 \cdot 0.58 \cdot 0.42 \\ &= 4159.019 \end{aligned}$$

Now, sample size must be a natural number, so we must round up any fractional-valued results. Common rounding which would often be rounding down will allow the margin of error to go slightly above the desired 1.5%, thus our need to always round up any resulting computed fractional amounts (the same is true for final interpretation of all sample size computations: we round up any fractional values.)

As a final summary, the researcher must collect a sample of at least size $4,160$ in order to keep the margin of error at most 1.5% while also requiring 95% confidence. If, upon reflection, the researcher decides it is unreasonable (possibly due to cost) to collect data from such a large number of eighth-grade students in Kansas, then either the allowed margin of error must be increased or else the confidence level must be decreased in order to decrease the required sample size.

2. A marketing researcher is interested in the proportion of European consumers who are aware of a U.S. branded product. The researcher is interested in controlling the margin of error to no more than 5% while working under a 99% confidence level. No previous study has been found about this topic. What size sample is required for the researcher to meet the desired conditions?

Answer

This time we proceed by applying our developed sample size formula in which we need the margin of error to be at most $\text{ME} = 5\%$ but in which we have no preliminary value of $\hat{p}$ known. So, again we must first determine the critical z-scores $z_{\frac{\alpha}{2}}$ tied to the prescribed confidence level of 99%. Therefore $\alpha = 1\%$ and $\frac{\alpha}{2} = 0.5\%$.

$$z_{0.005} = \pm\text{NORM.S.INV}(0.025) \approx \pm 2.57583.$$

Hence, our sample size calculation is

$$\begin{aligned} n &= \left(\frac{z_{\frac{\alpha}{2}}}{\text{ME}}\right)^2 \cdot 0.25 \\ &= \left(\frac{\pm 2.57583}{0.05}\right)^2 \cdot 0.25 \\ &= 663.4897 \end{aligned}$$

We once again round this results to a needed sample size of 664.

As a final summary, the researcher must conservatively collect a sample of at least size 664 in order to keep the margin of error at most 5% while also requiring 99% confidence.

It is worth noting that these are the minimum sample sizes needed if the sample is obtained via a simple random process. Other methods of sampling may require larger sample sizes. It is also worth noting that all the theory discussed in this section, as well as all the examples, operates under the assumption that the sample is a simple random sample, meaning, all samples of size n are equally likely. Use of the methodology developed here on samples not obtained in this way could lead to a much higher probability of inaccuracy. Bear this in mind when reading statistical analyses.

6.3: Confidence Intervals for Means (Sigma Known)

Learning Objectives

- Motivate the use of the z-score transformation to determine margin of error
- Define and compute critical values
- Determine the margin of error
- Construct confidence intervals, interpret their meaning, and apply them to contextual questions
- Utilize sample size as a means to balance confidence and margin of error

Confidence Intervals: A Quick Review

When we select a random sample and study it, we do not expect that the computed sample statistic is equal to the population parameter. The distance between the sample statistic and the population parameter is called the error. We want an idea of how far off our sample statistic might be from the population parameter and provide an interval of possible parameter values using the information from our sample. Through our knowledge of sampling distributions, we can provide a level of confidence that we have caught the population parameter in our interval. If the confidence level is 80%, the construction method successfully catches the population parameter for 80% of all the samples of that given size. In other words, if we repeatedly sampled the population randomly with the same sample size, we would expect 80% of the samples to produce confidence intervals with the population parameter in them. To maintain our level of confidence, we determine the distance (the margin of error), such that the percentage of samples have sample statistics that fall within that distance from the population parameter. If we then center our confidence interval at our computed sample statistic and extend our interval out by our margin of error in both directions, we produce a confidence interval that catches the population parameter with a success rate that is equal to our confidence level. We now dive into the details.

Confidence Intervals for Means

Let us frame our task within the particular context of this section: constructing confidence intervals for the population mean. We are constructing a confidence interval using information collected from a random sample of size n from our population. The form of our confidence interval will be $(\bar{x} - \text{ME}, \bar{x} + \text{ME})$, where $\bar{x}$ is the computed sample mean from the random sample of size n and ME is the margin of error. We must select a level of confidence for the confidence interval. This is the percentage of samples of size n that we want to be within the margin of error of the population mean. We need to determine ME, the margin of error, using the sampling distribution of sample means, which is normal, or at least approximately normal, under certain conditions. Those conditions must be met. If you cannot remember the conditions, you can review the section on the sampling distribution of sample means and commit the conditions to memory. Examine the figure below for a visual representation. Consider which of the symbols below will have a fixed, known value in an actual research situation.

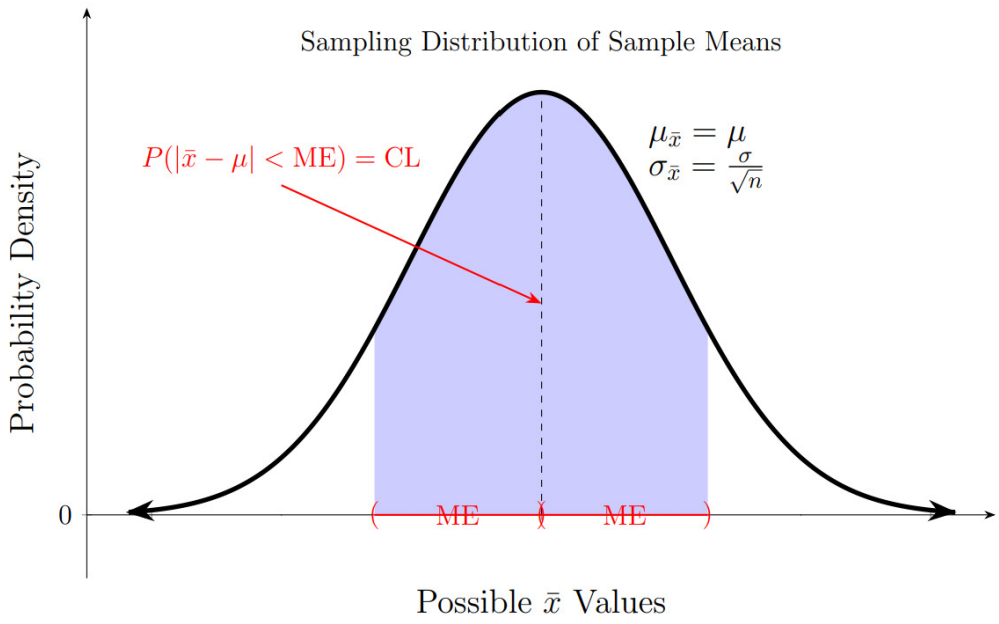


Figure 6.3.1: Sampling distribution of sample means

When considering the sampling distribution of sample means, $\bar{x}$ is a variable with no fixed value for us to know while determining the margin of error. Once we collect a sample and compute its sample mean, we will have a value of $\bar{x}$. It is important to remember that the logic of computing the margin of error requires us to treat $\bar{x}$ as a variable. We determine CL and n, so they are known to us. In general, we do not know anything about the population; that is why we are studying it. So, μ and σ are generally unknown as well. As such, it seems like our unknown symbols outnumber our known symbols. That is okay; we will be able to manage.

At this stage, we make one assumption for the sake of pedagogy. Let us assume that we know the value of σ, the population standard deviation. This is a rather large assumption because, as we all know, the population mean is an integral part of the computation of the population standard deviation. How could we know the population standard deviation without knowing the population mean? Perhaps in some situations, a past known population standard deviation may make a sufficient approximation for a current population standard deviation, but making such a claim is highly context dependent and beyond the scope of this book. For now, know we are making a simplifying assumption so that we can better understand the notion and construction of confidence intervals.

Determining the Margin of Error (σ known)

Recall that every normal distribution can be transformed into the standard normal distribution using the z-score transformation which preserves the area bounded beneath the probability density curves. We will use this transformation to determine what our margin of error needs to be. Consider the following sequence of figures that illustrates the transformation process.

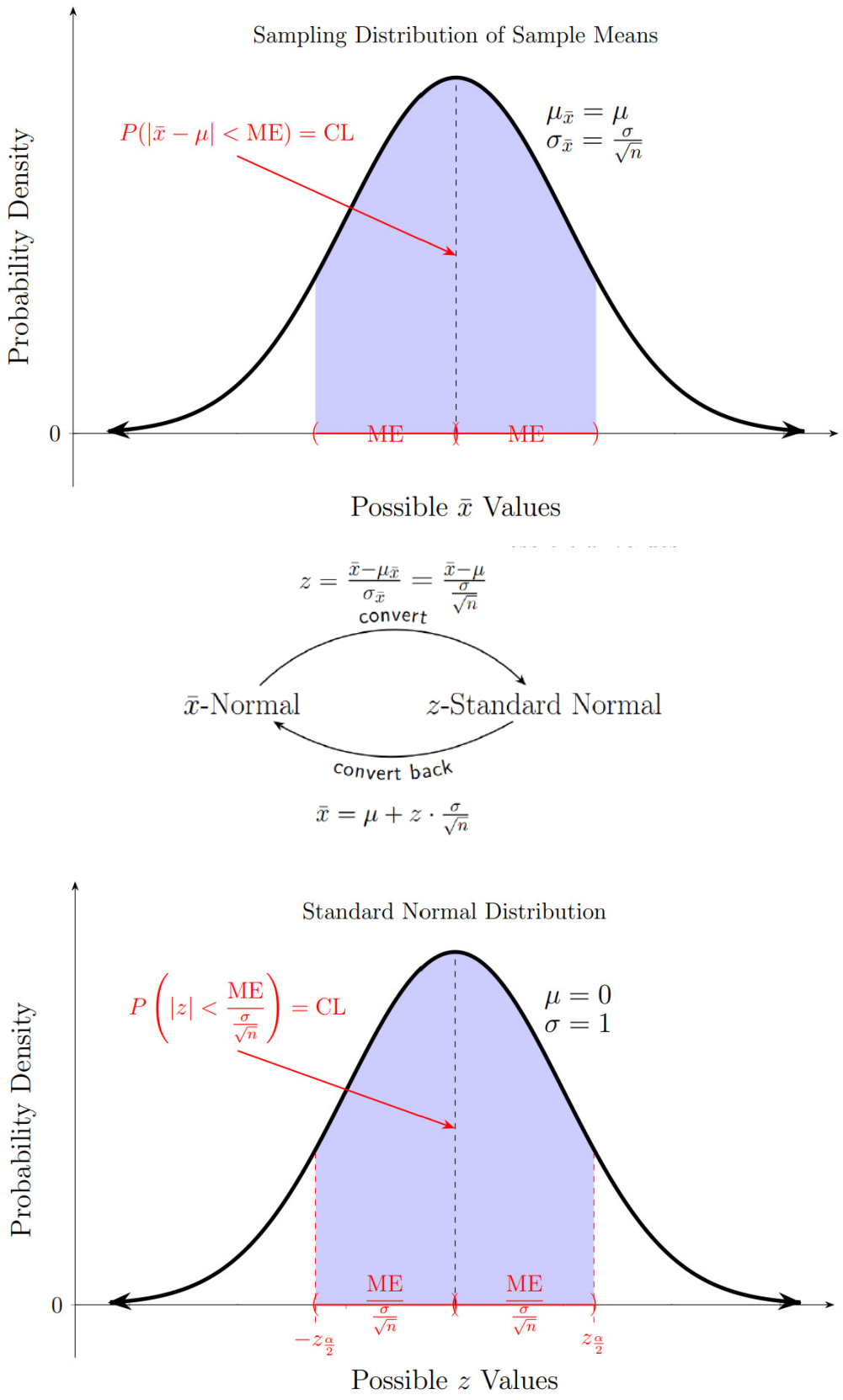


Figure 6.3.2: Sampling distribution of sample means transformed into the standard normal distribution

We started with the same figure as before, underwent the z-score transformation, and now want to determine the margin of error necessary to get the desired level of confidence. Notice that under the z-score transformation, the margin of error is scaled by a factor of one over the standard deviation of the sampling distribution. Since we are now considering the standard normal distribution, we know the mean and standard deviation of the distribution. Therefore, using technology, we can find the boundary points that will produce the desired confidence. These are denoted as $\pm z_{\frac{\alpha}{2}}$ in the figure above. We call these points **critical values**. Note that $z_{\frac{\alpha}{2}}$ represents the z-value where the area under the standard normal distribution to the right is $\frac{\alpha}{2}$ and (-z_{\frac{\alpha}{2}}\) represents the z-value where the area under the standard normal distribution to the left is $\frac{\alpha}{2}$.

? Text Exercise 6.3.1

Remaining in the context of constructing confidence intervals for population means when σ is known, determine the critical values for the indicated level of confidence by first sketching the problem in a standard normal distribution and then using technology to compute the critical values.

1. Confidence level: 90%

Answer

We first sketch a standard normal curve and then form an interval that is centered at the mean 0 and label the boundary points. The area under the curve between these two points is our confidence level. Notice that the critical values are equal in magnitude but opposite in sign so we can find one value and then take the positive and negative values as our critical values. Since we are using technology, we need to find the area to the left of one of the points. There are several ways of achieving this goal. We illustrate a different way for each of the first three problems of this text exercise; though, all three methods work for each of the problems.

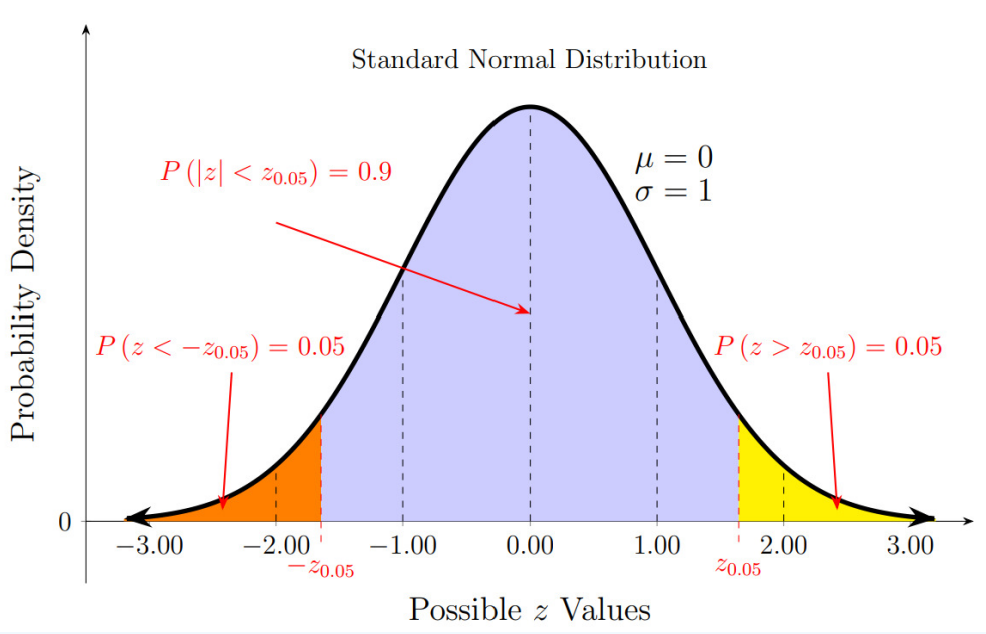


Figure 6.3.3 Standard normal distribution with 90% confidence interval

For our first example, we find the negative critical value first. We can determine the area outside of our critical values because the total area underneath the curve is 1, and the area between our critical values is 0.9. The area outside of our critical values is $1 - 0.9 = 0.1$. Note that this is what we have been calling the α value. Since normal distributions are symmetric about the mean and the critical values are equally far from the mean, the two tails of the distribution (the values less than the negative critical value and then the values greater than the positive critical value) have the same area. To find the area to the left of the negative critical value, we split the area of the two tails in half. $\frac{0.1}{2} = 0.05$. Notice the labels for the critical values; we replaced the $\frac{\alpha}{2}$ in the subscript with the value of $\frac{\alpha}{2}$ in the context of the problem. We can use technology to determine the left critical value.

$$-z_{0.05} = \text{NORM.S.INV}(0.05) \approx -1.6449$$

We thus have our critical values: $\pm z_{0.05} \approx \pm 1.6449$

2. Confidence level: 95%

Answer

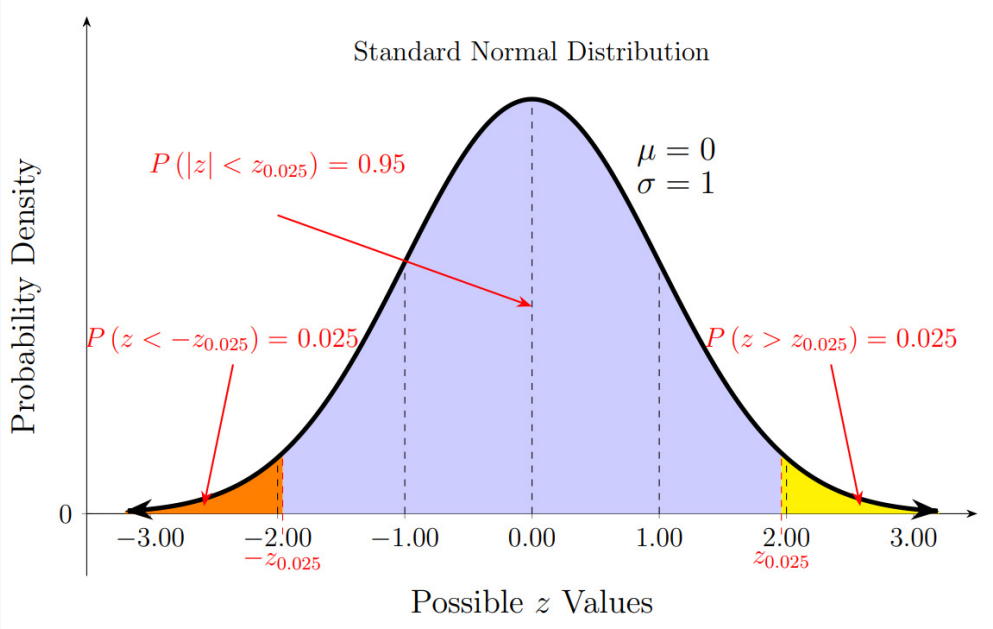


Figure $6.3.4$ Standard normal distribution with 95% confidence interval

We find the positive critical value for our second and third examples. Note that the area to the left of the positive critical value is the confidence level 0.95 and the area in the left tail, which we know from the last exercise is half of the α value $\frac{0.05}{2} = 0.025$. The area to the left of the positive critical value is $0.95 + 0.025 = 0.975$. We can use technology to determine the right critical value.

$$z_{0.025} = \text{NORM.S.INV}(0.975) \approx 1.96$$

Now we have our critical values: $\pm z_{0.025} \approx \pm 1.96$

3. Confidence level: 99%

Answer

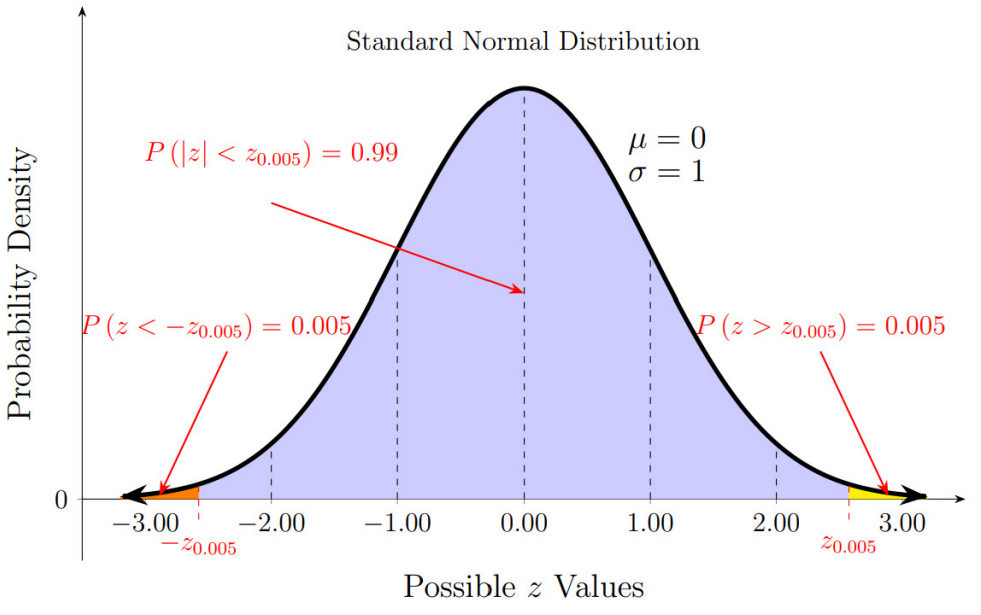


Figure 6.3.5 Standard normal distribution with 99% confidence interval

Another way to find the area to the left of the positive critical value is to use the complementary relationship between the area to the left and the area to the right of a point. The total area is 1. The area to the right of the positive critical value is half of the α value $\frac{0.01}{2} = 0.005$. So the area to the left of the positive critical value is $1 - 0.005 = 0.995$. We can use technology to determine the right critical value.

$$z_{0.005} = \text{NORM.S.INV}(0.995) \approx 2.5758$$

Now have our critical values: $\pm z_{0.005} \approx \pm 2.5758$

4. For a general confidence level CL

Answer

We generally care about the positive critical value as it represents the number of standard deviations that must be traversed in both directions from the mean to gain the desired confidence level. In this solution, we will use each of the three methods above to find the right critical value for the general confidence level CL.

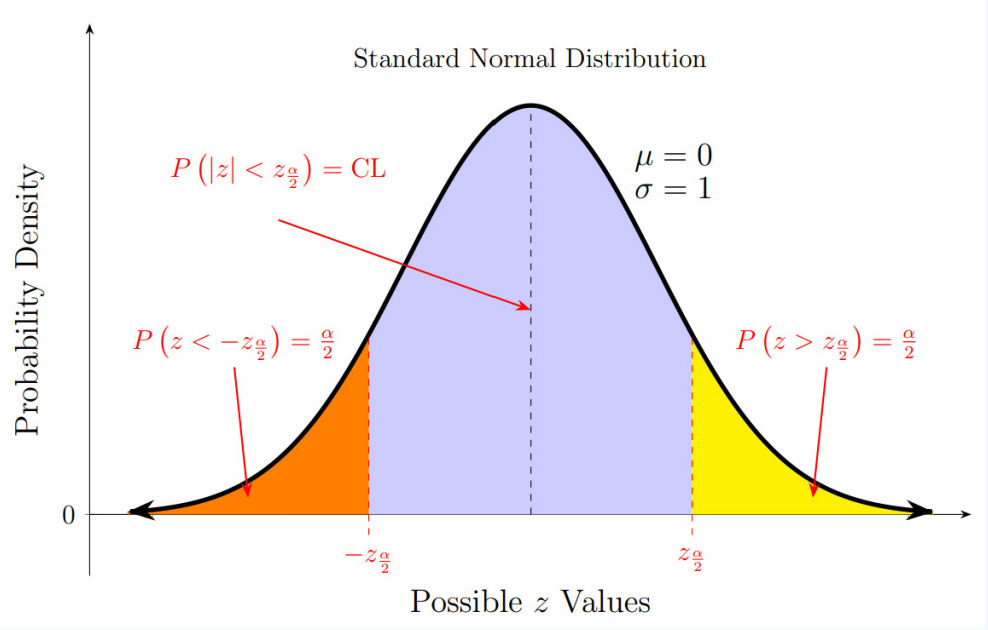


Figure 6.3.6 Standard normal distrubution with general confidence interval

$$\begin{aligned} z_{\frac{\alpha}{2}} &= -1 \cdot \text{NORM.INV}\left(\frac{\alpha}{2}\right) \\ &= \text{NORM.INV}\left(\text{CL} + \frac{\alpha}{2}\right) \\ &= \text{NORM.INV}\left(1 - \frac{\alpha}{2}\right) \end{aligned}$$

With the critical values in hand, we make the final step by noticing that the positive critical value is equal to the length of the scaled margin of error.

$$\frac{\text{ME}}{\frac{\sigma}{\sqrt{n}}} = z_{\frac{\alpha}{2}}$$

Given our simplifying assumption (that we know the population standard deviation), we know the factor by which the margin of error was scaled and what the scaled length is. From this, we determine the margin of error.

$$\text{ME} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Constructing Confidence Intervals for Means (σ known)

We now have all the pieces to construct a confidence interval for the population mean when the population standard deviation is known.

$$(\bar{x} - \text{ME}, \bar{x} + \text{ME}) = \left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

We often write these confidence intervals as $\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

? Text Exercise 6.3.2

The 2024 Toyota Camry Hybrid LE gets 52[1] miles per gallon when considering both highway and city driving with a standard deviation of 5.1 miles per gallon. In designing the 2025 Toyota Camry, the engineers would like to assert that the fuel efficiency in the newest model exceeds that of the previous model. The engineers randomly test-drove 50 2025 models and

recorded an average of 54 miles per gallon from the sample. Assuming the standard deviation remained the same, construct a 98% confidence interval to predict the population mean fuel economy. Does this bode well for the engineer's desires? Explain.

[1] This is the only statistic based on actual, substantial data. The remainder of the numbers in this problem were contrived loosely based on available data.

Answer

First, we check that the conditions for constructing a confidence interval for means are satisfied. We want our sample to be randomly selected and the sampling distribution to be approximately normal. Since the engineers randomly test drove 50 cars, we have both conditions met $(n > 30)$.

We next connect the values in the problem statement with the variables at play: $\text{CL} = 98\%$, $n = 50$, $\bar{x} = 54$, and $\sigma = 5.1$. Sometimes, we do not use all the numbers in a problem statement. 52 comes into play at the end, not while constructing the confidence interval, because the engineers want the population mean of the 2025 Camry to be greater than the previous model, which was 52 miles per gallon.

We need to find the positive critical value $z_{\frac{\alpha}{2}}$. Since $\text{CL} = 98\% = 0.98$, $\alpha = 1 - 0.98 = 0.02$, and $\frac{\alpha}{2} = \frac{0.02}{2} = 0.01$. The distribution that we find the critical value from is the standard normal distribution because we are considering means with the population standard deviation known. We encourage sketching pictures.

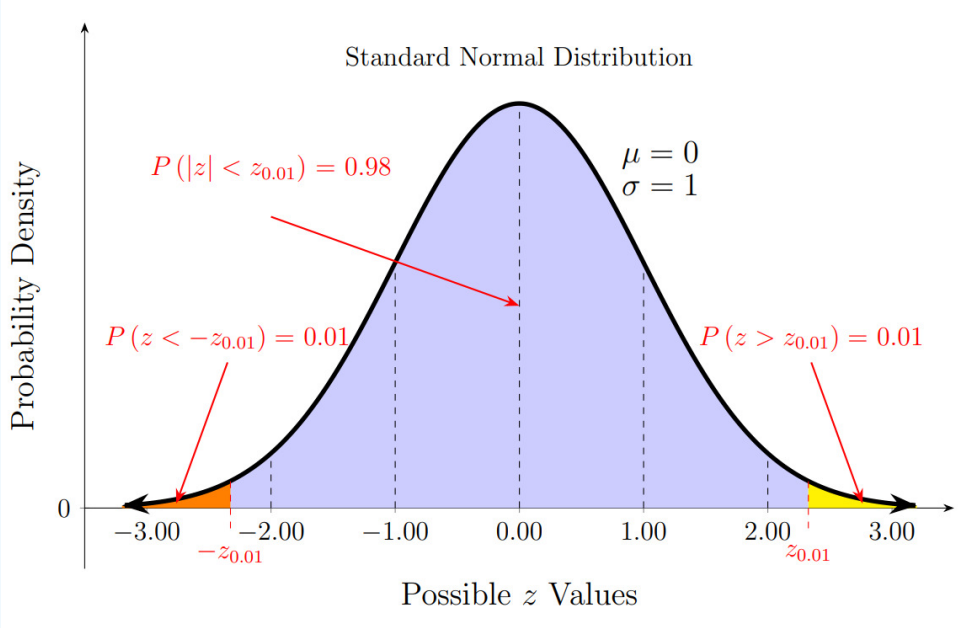


Figure 6.3.7: Standard normal distribution with 98% confidence interval

$$z_{0.01} = -1 \cdot \text{NORM.S.INV}(0.01) \approx 2.3264$$

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) \approx \left(54 - 2.3264 \cdot \frac{5.1}{\sqrt{50}}, 54 + 2.3264 \cdot \frac{5.1}{\sqrt{50}}\right) \approx (52.3221, 55.6779)$$

So, we have constructed the confidence interval based on the results of the random sample of 50 cars. Our conclusion is that, at a 98% confidence level, the population mean, μ, of the 2025 Camry Hybrid LE fuel economy is somewhere between 52.3221 miles per gallon and 55.6779 miles per gallon. We then notice that 52.3221 miles per gallon is greater than 52 miles per gallon. The engineers can feel confident that the fuel economy of the newest model exceeds the fuel economy of the previous model.

The Margin of Error and Sample Size

At the beginning of this section, we mentioned a balancing act at play in constructing confidence intervals. As we saw, the higher the confidence level, the larger the positive critical value. The larger the critical value, the larger the margin of error. At the same

time, we want our confidence interval to give us a pretty good idea of the population mean. The larger the margin of error, the wider the range of values we conclude our population mean falls. For example, we can be 100% confident that the population parameter falls in the interval $(-\infty, \infty)$, but that interval does not yield any useful information; similarly, we could very precisely estimate that $\mu = \bar{x},$ but we would have 0% confidence in this estimate. These desires conflict, but there is another variable at play in determining the margin of error: the sample size, n. We do not have control over σ, the population standard deviation, but we do have control over the size of the sample we select.

$$\text{ME} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Text Exercise 6.3.3

1. Explain what happens to the margin of error as the sample size n increases.

Answer

Since $\text{ME} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},$ we have that n is in the denominator. As n increases, $\sqrt{n}$ also increases. When dividing by larger and larger numbers, the resulting number is smaller and smaller. Consider $\frac{1}{1} = 1,$ $\frac{1}{10} = 0.1,$ $\frac{1}{100} = 0.01,$ and so on. As we increase the denominator by a factor of 10 each time, the value decreases getting closer and closer to 0. Thus, the margin of error goes to 0 as the sample size gets larger. This should match our intuition that larger samples are more likely to produce statistics close to the population parameter.

2. If the margin of error for a 95% confidence interval for means with σ known was 4 with a sample size of $35,$ how large of a sample must be taken to have a margin of error of 1 while maintaining the same level of confidence?

Answer

Since both confidence intervals are being constructed at the same level of confidence and from the same population, $z_{\frac{\alpha}{2}}$ and σ will be the same. We will have two margins of error and two sample sizes. $\text{ME}_1 = 4,$ $n_1 = 35,$ $\text{ME}_2 = 1,$ and n_2. The last one is unknown. This yields the following system of equations.

$$4 = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{35}}$$
$$1 = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n_2}}$$

We are only interested in finding n_2. So, we want to eliminate the critical value and standard deviation. We note that if we multiply the second equation by 4 on both sides, we can set the two equations equal to each other.

$$z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{35}} = 4 \cdot z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n_2}}$$
$$\cancel{z_{\frac{\alpha}{2}}} \cdot \frac{\cancel{\sigma}}{\sqrt{35}} = 4 \cdot \cancel{z_{\frac{\alpha}{2}}} \cdot \frac{\cancel{\sigma}}{\sqrt{n_2}}$$
$$\frac{1}{\sqrt{35}} = \frac{4}{\sqrt{n_2}}$$
$$\sqrt{n_2} = 4\sqrt{35}$$
$$\left(\sqrt{n_2}\right)^2 = \left(4\sqrt{35}\right)^2$$
$$n_2 = 16 \cdot 35 = 560$$

Text Exercise 6.3.4

1. Suppose the engineers at Toyota decided that they wanted a confidence interval with a margin of error of 0.25 miles per gallon while maintaining the confidence level of 98%, how large of a sample of 2025 Toyota Camry LE cars would need to be taken?

Answer

Recall that the population standard deviation was given to be 5.1 miles per gallon and that the positive critical value was $z_{\frac{\alpha}{2}} = -1 \cdot \text{NORM.S.INV}(0.01) \approx 2.3264$. The engineers have set the desired margin of error to $\text{ME} = 0.25$. Given the fact that $\text{ME} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, we can solve for the unknown sample size.

$$0.25 \approx 2.3264 \cdot \frac{5.1}{\sqrt{n}}$$
$$\sqrt{n} \approx \frac{2.3264 \cdot 5.1}{0.25}$$
$$(\sqrt{n})^2 \approx \left(\frac{2.3264 \cdot 5.1}{0.25} \right)^2$$
$$n \approx 2252.214$$

We now must remember the context behind the situation. We are trying to determine the minimum sample size necessary to result in a 98 confidence interval with a margin of error of 0.25 miles per gallon, and we have deduced that n must be at least 2252.214.We, therefore, decide that a sample of size 2253cars would be necessary.

2. Let us now solve the problem in general. If we set the margin of error, confidence level, and know the population standard deviation, how large of a sample is necessary to construct a confidence interval at that level of confidence and margin of error?

Answer

$$\text{ME} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$
$$\sqrt{n} = \frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\text{ME}}$$
$$(\sqrt{n})^2 = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\text{ME}} \right)^2$$
$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\text{ME}} \right)^2$$

Now, just as before, we need our sample size to be a whole number. When it is not a whole number, we always round up so that we are within the threshold of our margin of error tolerance. It is better to more precise than less precise.

Formula for the mathematically inclined

$$n = \left\lceil \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\text{ME}} \right)^2 \right\rceil$$

This formula introduces the ceiling function $\lceil x \rceil$ which returns the smallest integer value that is greater than or equal to x.

6.4: Confidence Interval for Means (Sigma Unknown)

Learning Objectives

- Form a basic intuition regarding the development and shape of t-distributions
- Introduce accumulation functions for t-distributions
- Discuss degrees of freedom
- Find critical values
- Construct confidence intervals for means using sample data
- Estimating necessary sample sizes for desired margin of errors

Section 6.4 Excel File (contains all of the data sets for this section)

Confidence Intervals for Means

Having developed a construction technique for confidence intervals for mean with σ known, we now drop our simplifying assumption and address the common case when we do not know much about the population; in particular, we do not know the population mean or the population standard deviation. Having an idea of what the sampling distribution of sample means looks like is paramount to our method. We must know that the sampling distribution of sample means is approximately normal. We encouraged the reader to review the section on sampling distributions of sample means in the last chapter; hopefully, the conditions are committed to memory, but we provide them now for quick reference: either the parent population is normal or the sample size n is greater than 30. Recall that this threshold works for many populations but not all.

The sampling distribution of sample means is approximately normal with a mean, $\mu_{\bar{x}} = \mu$, and a standard deviation, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Now, we do not know the value σ in addition to not knowing the value of μ. We set a level of confidence, which we understand as the percent of samples of size n that will produce a sample mean within a certain distance of the population mean; we call this distance the margin of error, ME. In the previous section, we determined the margin of error by transforming the sampling distribution of sample means into the standard normal distribution and finding our critical values. At this stage, we now run into difficulties because we do not know σ. We can approximate with the following transformation, which we call the t-transformation.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Notice, that we simply substituted σ with s, the sample standard deviation. Think about the ramifications of this; different samples will naturally produce different values for s. Thus, when we evaluate the t-transformation, we do not expect to get the same values that the z-score transformation would produce since that transformation used σ for every computation. We might expect that the distribution that the sampling distribution of sample means will not be normally distributed.

We can build the theory just as we did with sampling distributions: by examining populations where we have all the data, computing the value of the transformation for each sample, constructing a histogram, and identifying the general shape. The theory is formalized, just as with sampling distributions, with some sophisticated mathematics beyond the scope of this course, but we will, hopefully, build a basic intuition by trying to understand how the t-transformation compares to the z-score transformation.

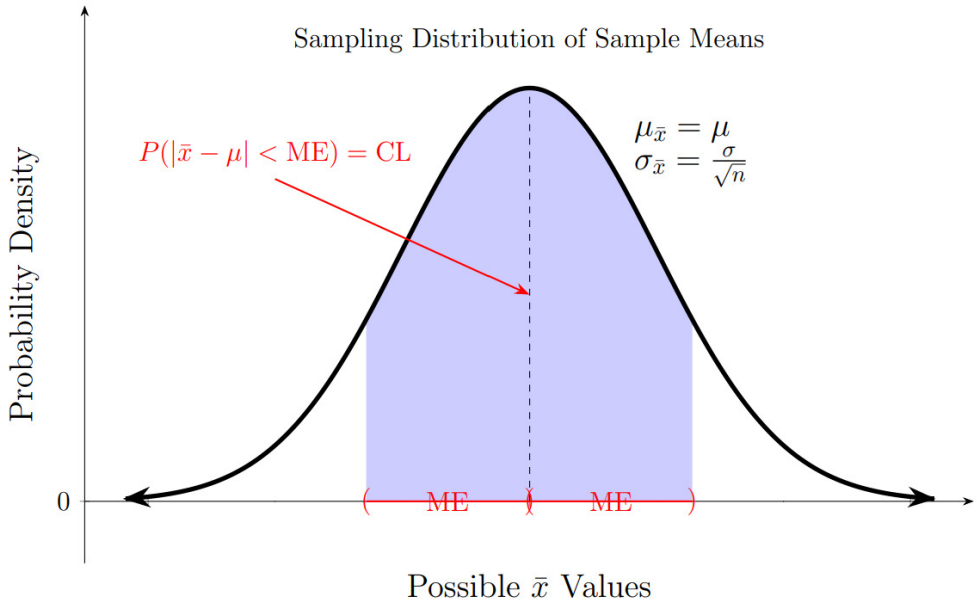
The sampling distribution of sample means is approximately normal, and normal distributions are symmetric, meaning, half of the samples of a given size produce sample means greater than μ and the other half of the samples produce sample means less than μ. We can also have samples with the same sample mean but with quite different standard deviations. So, we need to understand how these sample standard deviations are distributed. We are interested in the probability that our sample standard deviation is less than the population standard deviation. Note that the probability that the sample standard deviation is less than the population standard deviation is the same as the sample variance being less than the population variance. The sampling distribution of population variance is closely associated with the χ^2-distribution (an interested reader is encouraged to read or reread the sampling distribution of sample variances section for further details) which we have seen to be skewed right. Recall that when a distribution is skewed right, the mean is greater than the median and the probability that the sample variance is less than the population variance is greater than 50%. So, we are more likely to get sample standard deviations that are smaller than the population standard deviation than to get larger sample standard deviations.

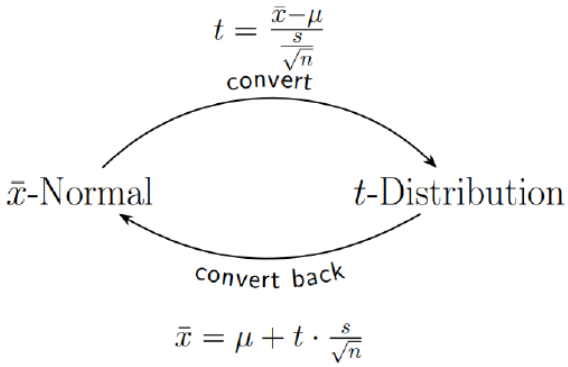
What does this all say about the t-transformation? Since we are just as likely to get sample means larger or smaller than the mean, we will be symmetric about 0. Since we are more likely to get sample standard deviations that are smaller than the population standard

deviation, we will usually be dividing by a smaller number in the t-transformation than in the z-score transformation. When dividing by smaller numbers, the quotient is larger. We expect larger magnitudes under the t-transformation than under the z-score transformation. This indicates that there is a greater probability density in the tails (the distribution has thicker tails). Hopefully, at this point, we recognize these descriptions as our descriptions of the Student's t-distribution. This t-distribution will have $n-1$ degrees of freedom, and we will explain why later in the section.

Determining the Margin of Error (σ unknown)

Now that we have built an intuitive understanding of the$t$-distribution, we are prepared to determine the necessary margin of error in the context of not knowing the population standard deviation. We provide a similar progression of figures to illustrate the process below.





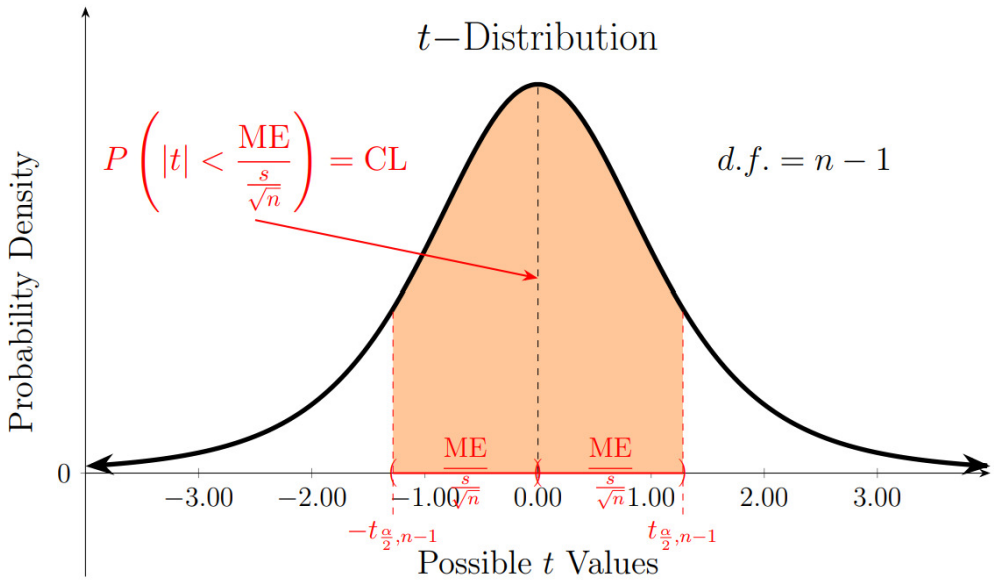


Figure 6.4.1: Sampling distribution of sample means under the t-transformation

Knowing that the sampling distribution of sample means is transformed into a t-distribution with $n-1$ degrees of freedom enables us to determine the necessary margin of error for our desired confidence level. Just as when σ is known, the margin of error is scaled by a known factor known to us. And once we determine the boundary points of our shaded region, the critical values $\pm t_{\frac{\alpha}{2},n-1}$, we can compute the margin of error. Notice the additional subscripts in the notation for our critical values of the t-distribution due to the critical values depending on the confidence level and the degrees of freedom. Once again, we must use technology to determine critical values with an accumulation functions specific to the t-distribution.

In Excel, we utilize the T.DIST and T.INV functions which work very similarly to the NORM.DIST and NORM.INV functions that we have been working with for quite some time. We use the distribution function to find the area to the left of a point. Using T.DIST, we enter the point that we want the area to the left of, and the necessary information to describe the distribution; for normal distributions, we used the mean and standard deviation, but for t-distributions, we use the degrees of freedom. Finally, we then tell the function to accumulate.

$$P(t < a) = \text{T.DIST}(a, d.f., 1)$$

The T.INV function is used to find the point in a t-distribution with a certain number of degrees of freedom such that a given area is to the left of the point. Using T.INV, we enter the area and the degrees of freedom.

$$a = \text{T.INV}(\text{area to the left of } a, d.f.) = \text{T.INV}(P(t < a), d.f.)$$

? Text Exercise 6.4.1

Remaining in the context of constructing confidence intervals for population means when σ is unknown, determine the positive critical value for the 95% confidence level given the indicated sample size by roughly sketching the t-distribution and then using technology.

The rough sketches drawn by hand are important to ensure a proper approach to the problems, but we cannot accurately depict what happens to the distributions as the degrees of freedom change. As you complete each part of this text exercise, examine the computer-generated graphics to solidify what happens as the degrees of freedom increase.

1. $n = 4$

Answer

We begin by sketching the t-distribution with 3 degrees of freedom. It is symmetric about 0 and has thicker tails than the standard normal distribution. We then form an interval centered at 0 and label the boundary points. The area under the curve between these two points is our confidence level. Notice that the critical values are equal in magnitude but opposite in sign.

Due to the symmetry of the distribution, the two tails have equal area giving $\frac{\alpha}{2}$ in each tail. Since our confidence level is 95%, the α value is 0.05meaning the area in each tail is 0.025.

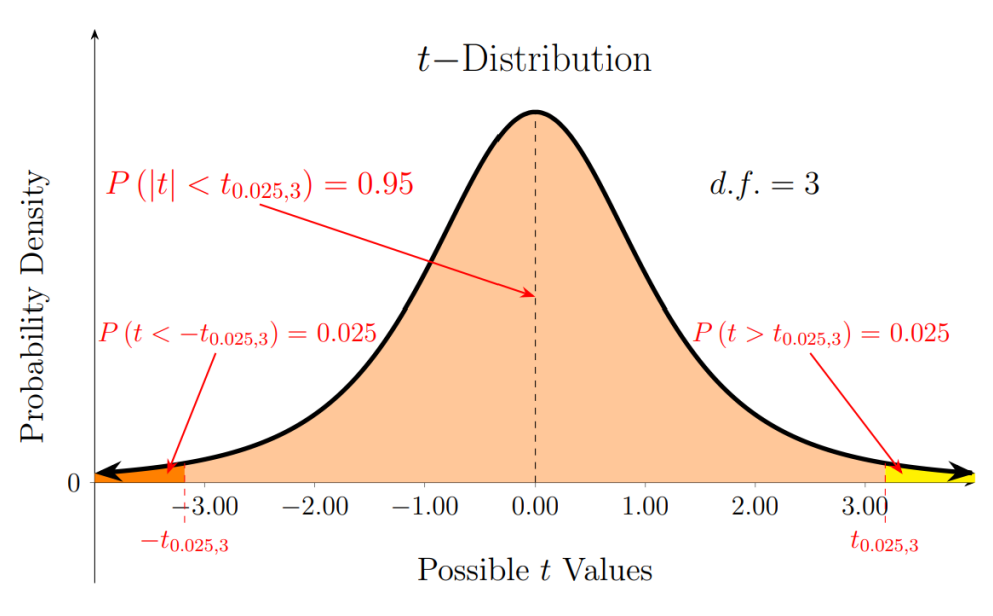


Figure $6.4.2$ t-Distribution with 3 degrees of freedom

Since the critical values are equal in magnitude but opposite in sign, we can compute the positive critical value by multiplying the negative critical value by -1.

$$t_{0.025,3} = -1 \cdot \text{T.INV}(0.025, 3) \approx 3.1825$$

2. $n = 8$

Answer

Keeping the same level of confidence but with a larger sample, our tasks throughout this text exercise are similar. The t-distribution now has 7 degrees of freedom, but the areas and the process remain the same.

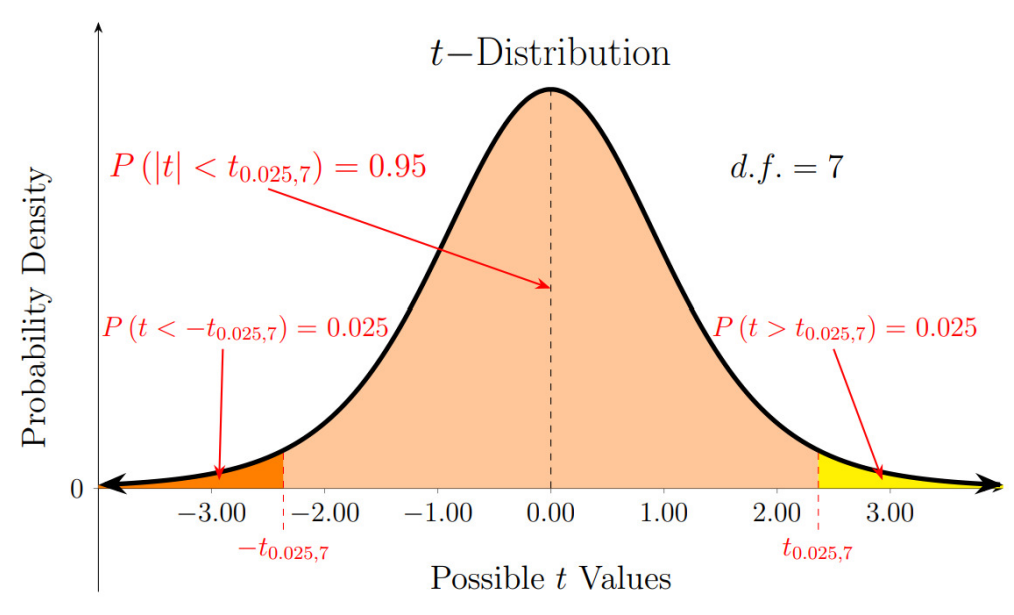


Figure $6.4.3$ t-Distribution with 7 degrees of freedom

$$t_{0.025,7} = -1 \cdot \text{T.INV}(0.025, 7) \approx 2.3646$$

3. $n = 16$

Answer

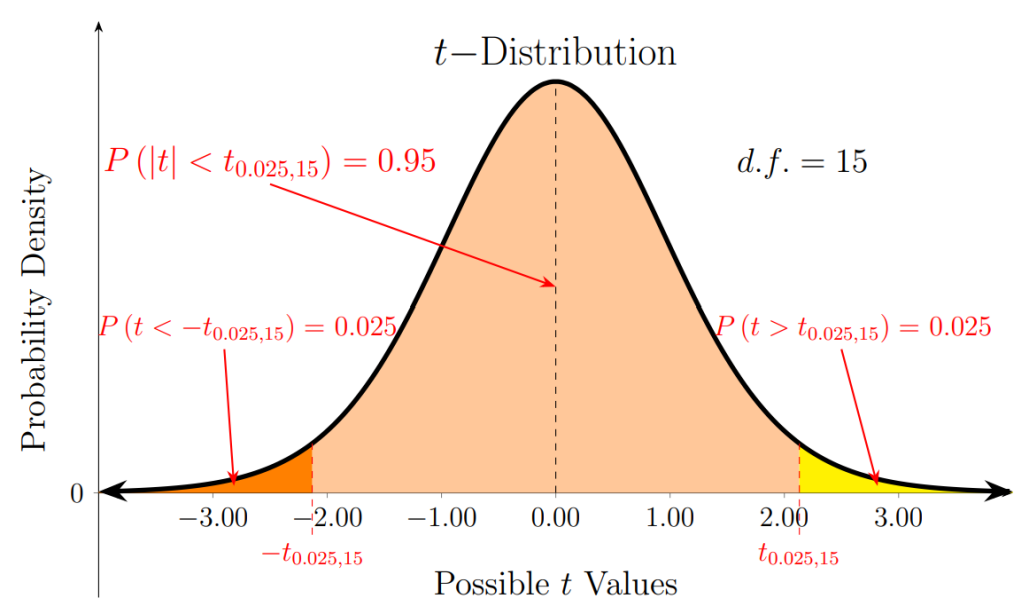


Figure 6.4.4 t-Distribution with 15 degrees of freedom

$$t_{0.025,15} = -1 \cdot \text{T.INV}(0.025, 15) \approx 2.1315$$

4. $n = 32$

Answer

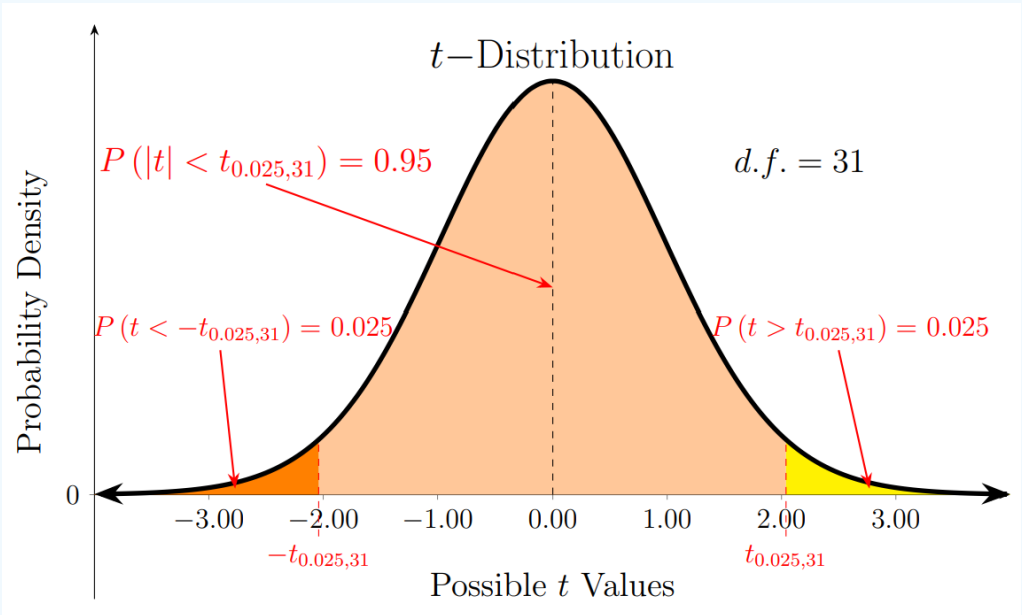


Figure 6.4.5 t-Distribution with 31 degrees of freedom

$$t_{0.025,31} = -1 \cdot \text{T.INV}(0.025, 31) \approx 2.0395$$

5. Describe what is happened to the t-distributions as the degrees of freedom increased. What happened to the positive critical values? What will happen if we use larger and larger sample sizes?

Answer

Since each figure is plotted with the same horizontal axis, we can tell that the area under the curve in the tails decreases fairly noticeably with each subsequent figure. Since the total area is always 1, there is more area in the central portion of the distribution. We can see this faintly as the curve rounds out near its peak and thickens ever so slightly by the labels of the areas in the tails. The critical values are decreasing in magnitude with each increase in the degrees of freedom. This makes sense because the tails are getting thinner. The rate at which the tails are thinning and the critical values are decreasing in magnitude is slowing down. Recall that when we first introduced the t-distribution, we said that the distribution gets closer and closer to the standard normal distribution as the degrees of freedom increase. As such, we can expect the critical values of the t-distributions to approach the critical value of the standard normal distribution $z_{0.025} \approx 1.96$ as the sample size increases. With 61 degrees of freedom, we finally are less than 2. Only after 473 degrees of freedom will answers rounded to two decimal places match for this level of confidence.

Note: Degrees of Freedom

Recall that the shape of the t-distribution is determined by a quantity called degrees of freedom $(d.f.)$. Notice that the degrees of freedom in the figure above are related to the sample size, n; in particular, $d.f. = n - 1$. We shall now discuss the meaning of the terminology. In reading "degrees of freedom", we might naturally think about the extent to which someone is capable of determining and carrying out an action. That is a fine initial intuition; along those lines, we can think of the degrees of freedom as a measure regarding the extent to which the data can vary independently, given any constraints on the data.

Suppose the sample mean of 20 observations was 10. With the current information that we have, we do not know anything about the actual values; all we know is that the sum of the 20 observations is 200. There are infinitely many possible values for our 20 observations that result in such a mean. If it was then revealed to us that the first observation value was 11, we would know the sum of the last 19 observation values is 189. There are again infinitely many possible values for these 19 observations. If it was then revealed that the second observation value was 15,, we would know the sum of the last 18 observation values is 171, which can again happen in infinitely many ways. At what point, after revealing many observation values, will we know what the remaining observation values have to be knowing the sample mean is 10? Once 19 of the observation values are revealed, we will be able to deduce the last observation without it being revealed. We thus say that the first 19 observations are free to vary independently, but the 20^{th} observation value depends entirely on the previous observation values since the sample mean is known. Hence, there are 19 $(n-1)$ degrees of freedom.

When collecting data via random sampling, each observation value is independent of the others; the information gleaned from each observation provides information independent of the other observations. However, once we know, say, the sample mean, a constraint is placed on the data. The observation values are now connected through the expression $x_1 + x_2 + \ldots + x_n = \bar{x} = 20$. Each bit of information is no longer perfectly independent. Degrees of freedom is a measure for how much independent information remains given the constraint(s) in place.

So, why does the t-distribution have $n-1$ degrees of freedom in this situation? The sample mean only requires the observation values; there are no constraints connecting the data within the definition of the sample mean. This is not the case when we compute sample standard deviation. Here, we consider all the square deviations from the mean as the fundamental pieces of information, but implicit within this description is an expression that connects all data values: the sample mean. Each square deviation is not perfectly independent from every other square deviation. Since we needed the sample mean to compute the sample standard deviation, we have a constraint within the analysis. As we have seen above, this constraint reduces the amount of independent information by 1.

As a general rule of thumb, we expect the degrees of freedom to be the sample size minus the number of statistics used within computations of the statistics involved in the analysis. Different statistical analyses have different degrees of freedom; so, it is important to understand where the measurement comes from and to pay close attention to the literature explaining any particular statistical analysis.

With the critical values in hand, we again notice that the positive critical value is equal to the length of the scaled margin of error and determine our margin of error from there.

$$\frac{\text{ME}}{\frac{s}{\sqrt{n}}} = t_{\frac{\alpha}{2},n-1}$$

$$\text{ME} = t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$$

Constructing Confidence Intervals for Means (σ unknown)

We now have all the pieces to construct a confidence interval for the population mean.

$$(\bar{x} - \text{ME}, \bar{x} + \text{ME}) = \left(\bar{x} - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

We often write these confidence intervals as $\bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$.

? Text Exercise 6.4.2

We have discussed the heights of adult females on several occasions; they are normally distributed with a mean of 64 inches with a standard deviation of 2.5 inches. We have yet to discuss the heights of adult males. It seems reasonable to think that if the heights of adult females are normally distributed, the heights of adult males will also be normally distributed (this is indeed the case). We want to build a 90% confidence interval to catch the average height of adult males. To do so, we randomly sampled 10 adult males. Their heights in inches from shortest to tallest are reported below. Construct and interpret the confidence interval.

$$64, 66, 67.5, 68, 69, 69.75, 71.25, 72, 73, 73.25$$

Answer

In order to construct a confidence interval for means, we need the sampling distribution of sample means to be at least approximately normal and the sample to have been randomly selected. Given that the heights of adult males are normally distributed, we have that the sampling distribution of sample means is normally distributed despite the fact that we only have a sample size of 10. The sample was randomly chosen; we can, therefore, proceed.

We need the sample mean, standard deviation, sample size, and positive critical value to construct the confidence interval. We report, without explanation, that $\bar{x} = 69.375$ inches, $s \approx 3.0647$ inches, and $n = 10$. To calculate the positive critical value, we must find α. Since $\text{CL} = 0.90, \alpha = 0.1$. Since we do not know the population standard deviation, our critical value comes from the t-distribution with 9 degrees of freedom. We sketch the distribution to help us compute the positive critical value.

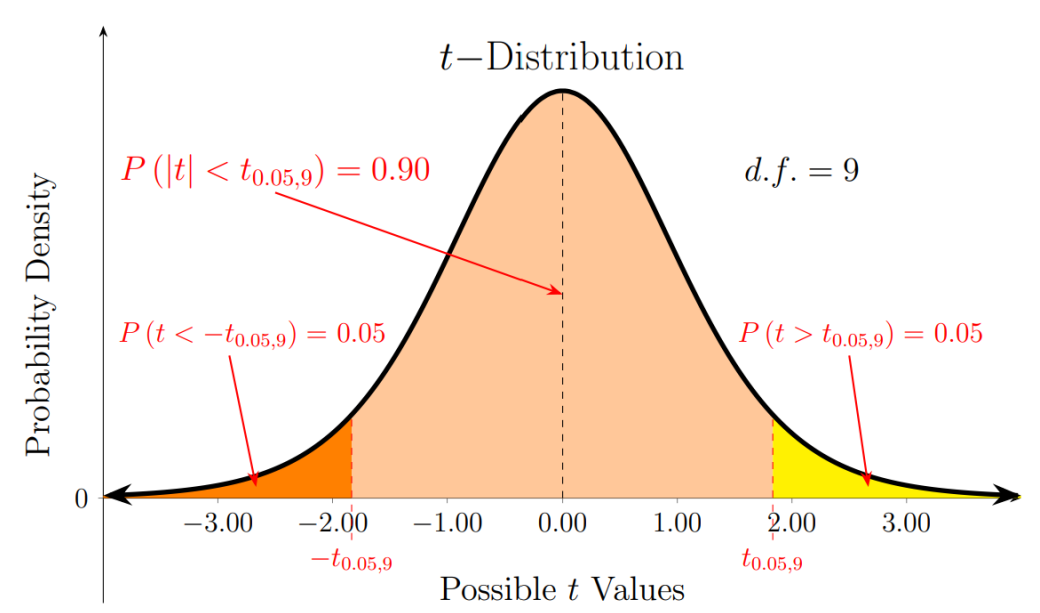


Figure 6.4.6 t-Distribution with 9 degrees of freedom

$$t_{0.05,9} = -1 \cdot \text{T.INV}(0.05, 9) \approx 1.8331$$

$$\left(\bar{x} - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}\right) \approx \left(69.375 - 1.8831 \cdot \frac{3.0647}{\sqrt{10}}, 69.375 + 1.8831 \cdot \frac{3.0647}{\sqrt{10}}\right) \approx (67.5985, 71.1516)$$

At a confidence level of 90%, the average height of all adult males μ is between 67.5985 inches and 71.1516 inches.

For many, the confidence interval produced in the previous text exercise is rather disappointing; there is a wide range of values that the population mean could be. We might desire to determine how large of a sample would be sufficient to expect that the margin of error is less than some specified value. In our case, we might be interested in determining how large of a sample would be sufficient to expect that at most one whole number falls in the constructed confidence interval while keeping the confidence level at 90%.

A natural place to begin would be with the margin of error formula and trying to solve for n, just as we did in the last section.

$$\text{ME} = t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$$

Difficulties, however, arise in several places. There are two values in the equation that depend on the value of n, the critical value and the square root of n. We also do not know what our sample standard deviation will be without actually collecting the sample. Unfortunately, these issues cannot be remedied perfectly, but there are paths forward that can arrive at reasonable estimates of a sufficient sample size. These estimates on n are just that, estimates; they will not be guaranteed to work without fail, but in general, we can rely on them to continue forward. The reader is encouraged to take more advanced statistics courses for a thorough explanation. We shall only provide some basic intuition about the considerations. We first solve for the sample size n.

$$n = \left(\frac{t_{\frac{\alpha}{2},n-1} \cdot s}{\text{ME}}\right)^2$$

We have not solved for n explicitly because the critical value from the t-distribution depends on n. Our path forward in estimating n is to replace the values in the numerator with values that we believe are larger than the values that will end up being used when we actually construct the confidence interval. When we replace one value in the numerator with one that is larger in the expression, the product will be larger. If we use this larger product as our estimated n,, we have probably chosen a larger than necessary sample size for our desired precision. Doing this will give us a conservative estimate of the sample size. We could always be wrong about whether the values were larger or not and then possibly not reach the desired precision in the confidence interval. Let us go through this value by value.

How do we pick a large enough value to overestimate the critical value? Recall that as the degrees of freedom increase, the t-distributions get closer and closer to the standard normal distribution and that the t-distribution has fatter tails than the standard normal distribution. These two facts imply that for a given confidence level, the critical values of the t-distributions are farther away from 0 than the critical values of the standard normal distribution, but as n increases the critical values of the t-distributions approach the critical values of the standard normal distribution. This means that the critical values in t-distributions get smaller in magnitude as n increases. This gives a conservative overestimate of the critical value by using the positive critical value at the same level of confidence from a t-distribution with fewer degrees of freedom than you expect to have from your future sample. If the underlying distribution is not normal, we can expect to need at least a sample size of 31, so $t_{\frac{\alpha}{2},30}$ would be an overestimate because we know we need at least 31 observations in our sample. If the underlying distribution is normal, we could have less in our sample. In this case, we could use the t-distribution with only 1 degree of freedom. It will be at least as large as every possible critical value. This is the most conservative estimate we can make.

Just as with constructing confidence intervals, there is a balancing act at play, not between confidence and precision, but between confidence and work. The more conservative our estimate, the larger the estimated sample which requires more work on the part of the researcher. There are time and financial constraints involved. Sometimes, researchers are satisfied with less conservative estimates which is fine since the most conservative estimates are overestimates by a large margin. Indeed, some textbooks recommend using the critical value from the standard normal distribution, which is guaranteed to be an underestimate of the critical value used when constructing the confidence interval.

The decision to choose a large enough value to overestimate the sample standard deviation is a more difficult question. In essence, the sample standard deviation is an estimate for the population standard deviation. We know that the sample standard deviation tends to be an underestimate of the population standard deviation but that it can be larger. We have computed a sample standard deviation from an initial study, but we do not know if it is larger than the population standard deviation or smaller. We expect that for larger samples our

computed standard deviation is closer to the population standard deviation but that is not necessarily the case. As we might see (if you study all of the bonus material, you will), it is sometimes possible to construct a confidence interval for the population standard deviation. With such an estimate, we could estimate the upper bound of where we are confident the population standard deviation falls, but even that could fail us. As such, some researchers construct confidence intervals from the pilot data. Others just use the sample standard deviation found in the sample data. And, yet others add a certain amount to the computed sample standard deviation. Here is where our guarantee must fail, but again that does not mean the estimate is not useful for continuing. We will be satisfied using the sample standard deviation from the pilot study in this text.

Within this text, we adopt the practice of using the pilot study sample standard deviation, comparing both $t_{\frac{\alpha}{2},1}$ and $t_{\frac{\alpha}{2},30}$ when the underlying distribution is normal, and otherwise just using $t_{\frac{\alpha}{2},30}$ in our estimates. In this last case, if our estimated value is less than 31, we select 31 instead. Let us now estimate the sample size such that we expect to have only one whole number in confidence interval.

We have a sample of size 10 and computed its sample standard deviation. We do not expect this standard deviation to be equal to the population standard deviation or the future sample to have the same standard deviation as this sample, but we will use $s = 3.0647$ to approximate what the sample standard deviation of a future sample might be. Since the underlying distribution is normal, we will compare $t_{0.05,1} \approx 6.3138$ and $t_{0.05,30} \approx 1.6973$ as overestimates of the critical value.

We also need to determine the margin of error specified in the problem. If there is to be at most one whole number in the interval, the interval length must be less than 1 for the distance between consecutive whole numbers is 1. Thus if the length of our interval is larger than 1, it is possible to have two whole numbers in our confidence interval. If the length is less than or equal to 1, then it is possible that we do not have any whole numbers in our interval, but that is permitted in the phrasing of the question. Thus the maximal length of the interval so that we have at most one whole number in the interval is 1. Since the margin of error is half of the interval length, our desired margin of error is confirmed to be 0.5 inches. We thus consider the following two cases.

$$n \leq \left(\frac{t_{0.05,1} \cdot s}{\text{ME}} \right)^2 \approx \left(\frac{6.3138 \cdot 3.0647}{0.5} \right)^2 \approx 1497.652$$

$$n \leq \left(\frac{t_{0.05,30} \cdot s}{\text{ME}} \right)^2 \approx \left(\frac{1.6973 \cdot 3.0647}{0.5} \right)^2 \approx 108.2264$$

In the first case, the estimated sample size is 1498; in the second case, the estimated sample size is 109. In both cases, the estimated sample size is larger than 30. As such, we use the estimate from case 2. We expect that a random sample of 109 adult males will produce a confidence interval that contains at most 1 whole number.

? Text Exercise 6.4.3

A pilot study of 9 randomly selected college students revealed that the average time spent scrolling on social media per day was 2.2 hours with a standard deviation of 2.1 hours. The researchers plan to conduct a larger study to construct a confidence interval at a confidence level of 99 with a margin of error of 1.25 hours. Estimate the number of college students that should be randomly sampled in the larger study to produce the desired results.

Answer

The goal is to build a confidence interval at the 99% confidence level with a margin of error that is no more than 1.25 hours. The sample standard deviation from the pilot study is 2.1 hours. We do not know anything specific about the population. We, therefore, use $t_{0.005,30} \approx 2.75$ to replace our critical value in the estimation.

$$n \leq \left(\frac{t_{0.005,30} \cdot s}{\text{ME}} \right)^2 \approx \left(\frac{2.75 \cdot 2.1}{1.25} \right)^2 \approx 21.3443$$

Using the critical value for a sample size of 31 produced an estimated sample size of 22. We cannot use this result for two reasons, in order to build our confidence interval in this situation, we need a sample size of at least 31. The critical value at the 99% confidence level with a sample of 22 will be larger than the critical value used in our estimate; we, therefore, must proceed with caution. We are not guaranteed that the value would not be sufficient given the possibility that the sample standard deviation in the large study might be smaller than in the pilot study. Luckily for us, the decision was forced with the original consideration. If the underlying distribution is not normal and we do not know any more information about its

distribution, we expect the sampling distribution of sample means to be approximately normal when $n > 30$ which is an important requisite for constructing confidence intervals.

6.5: Confidence Intervals for Variances - Optional Material

Learning Objectives

- Develop a second general methodology of constructing confidence intervals
- Find critical values in the χ^2-distribution
- Construct confidence intervals for variances and standard deviations using sample data

Confidence Intervals for Variances

At this point in our development of confidence intervals, we introduce another methodology. The general interpretation of confidence intervals remains the same: a confidence interval built at an 85% confidence level catches the population parameter for 85% of all samples of that given size, or alternatively, catches the population parameter for 85% of the confidence intervals constructed if random sampling is repeatedly conducted. We, however, will no longer base the construction methodology around the idea that 85% of all sample statistics fall within the margin of error of the population parameter. This worked really well when the sampling distributions were approximately normal and hence symmetric about the population parameter, but the sampling distributions for variances are not symmetric. Recall that the sampling distributions of sample variances were not normal, but skewed right, and that we could transform them into χ^2-distributions to determine probabilities. We studied the sampling distribution of sample variances only when the parent population was normal; we shall remain in this realm of normal parent populations throughout this section.

Previously when we were constructing confidence intervals, we routinely produced a figure split into three regions with known areas within a specific distribution: the left tail, the right tail, and the central region. Each tail had an area that was equal to $\frac{\alpha}{2}$, and the central region had an area equal to CL. We shall begin our development with such a figure within the context of randomly drawing a sample of size n from a population normally distributed with the intent to build a confidence interval for the population variance at the confidence level CL. We sketch such regions in the following figure of the χ^2-distribution with $n-1$ degrees of freedom.

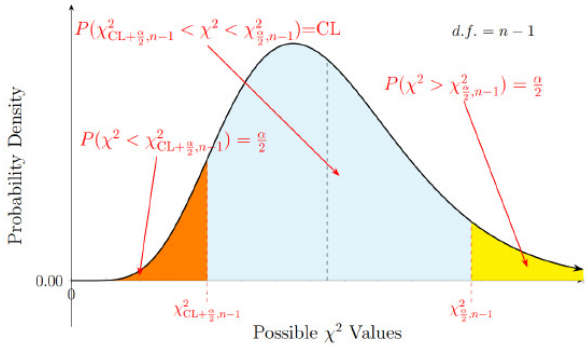


Figure $6.5.1$: χ^2-distribution

The boundary points of the central region are again called critical values; just like the critical values in the t-distribution, they depend on the confidence level and the degrees of freedom. There are, however, some significant differences. Notice that they are not the same distance away from the mean of the distribution (the black dashed line); the distribution is positively skewed, and we constructed the regions so that the tails each have an area of $\frac{\alpha}{2}$. Observe that both critical values are positive. To help us distinguish between the two critical values, we use the first part of the subscript to indicate the area to the left of the critical value. Note the smaller critical value has $\text{CL}+\frac{\alpha}{2}$ to the right of it; while, the larger critical value has $\frac{\alpha}{2}$. We label the critical values as seen in the figure. The smaller critical value is $\chi^2_{\text{CL}+\frac{\alpha}{2},n-1}$, and the larger critical value is $\chi^2_{\frac{\alpha}{2},n-1}$.

Now that we have an understanding of the figure and its labels, consider the probability statement at the top of the figure.

$$P\left(\chi^2_{\text{CL}+\frac{\alpha}{2},n-1} < \chi^2 < \chi^2_{\frac{\alpha}{2},n-1}\right) = \text{CL}$$

It says that the probability that the random variable χ^2 with $n-1$ degrees of freedom falls between the two critical values is equal to the confidence level. In our context, this χ^2 variable with $n-1$ degrees of freedom is related to the sampling distribution of

sample variances through the following transformation.

$$\chi^2_{n-1} = \frac{(n-1)}{\sigma^2} \cdot s^2$$

We can understand the probability statement about the random variable χ^2 in terms of the random variable s^2, the sample variance.

$$P\left(\chi^2_{\text{CL}+\frac{\alpha}{2},n-1} < \frac{(n-1)}{\sigma^2} \cdot s^2 < \chi^2_{\frac{\alpha}{2},n-1}\right) = \text{CL}$$

Recall that the underlying random experiment for the random variable s^2 is randomly selecting a sample of size n from a population that is normally distributed. We can think of this probability statement as follows: the probability of randomly selecting a sample of size n so that the sample variance scaled by $\frac{(n-1)}{\sigma^2}$ falls between the critical values is the confidence level. Recall that we are interested in constructing a confidence interval for σ^2. We can use algebraic manipulation to get σ^2 isolated in the expression.

$$P\left(\chi^2_{\text{CL}+\frac{\alpha}{2},n-1} < \frac{(n-1)}{\sigma^2} \cdot s^2 < \chi^2_{\frac{\alpha}{2},n-1}\right) = \text{CL}$$

$$P\left(\frac{\chi^2_{\text{CL}+\frac{\alpha}{2},n-1}}{(n-1)s^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{\frac{\alpha}{2},n-1}}{(n-1)s^2}\right) = \text{CL}$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{\text{CL}+\frac{\alpha}{2},n-1}} > \sigma^2 > \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right) = \text{CL}$$

This last step might require further explanation. We want an expression with σ^2 not with $\frac{1}{\sigma^2}$. Notice that these two expressions are reciprocals of each other. So, we reciprocate each term in the string of inequalities and figure out what happens with the inequalities. Consider the very simple string of inequalities $2 < 3 < 4$. The reciprocals of each of the terms in our inequality are $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4}$. Notice that $\frac{1}{2} > \frac{1}{3} > \frac{1}{4}$. We are in a similar situation in our probability statements. We have a string of inequalities with positive values in each term; so, when we reciprocate, we must flip the inequality signs. We generally have lower bounds on the left and upper bounds on the right; so, we reorder this last line to arrive at a final probability statement.

$$P\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\text{CL}+\frac{\alpha}{2},n-1}}\right) = \text{CL}$$

We understand this last probability statement to say that the probability of randomly selecting a sample of size n from the normally distributed parent population so that $\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}$ is less than σ^2 and $\frac{(n-1)s^2}{\chi^2_{\text{CL}+\frac{\alpha}{2},n-1}}$ is greater than σ^2 is the confidence level. In other words, we have constructed an interval so that the population variance falls within that interval the confidence level percent of the time. Thus, this construction of confidence intervals for variances yields confidence intervals of the following form.

$$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{\text{CL}+\frac{\alpha}{2},n-1}}\right)$$

Constructing Confidence Intervals for Variances

We now have a method of constructing confidence intervals for variances that produces confidence intervals of a specific form; this form is quite different from the forms for means and proportions. The sample statistic is no longer the center of the interval. Critical values still play an important role in the construction of the interval, and computing these critical values is the last aspect of construction that we need to hammer out. In the section on sampling distributions of sample variances, we introduced the CHISQ.DIST accumulation function. In order to calculate critical values, we need an inverse accumulation function. We introduce CHISQ.INV which works similarly to the inverse accumulation functions that have been previously introduced. Given an area and the degrees of freedom of the distribution, CHISQ.INV returns the point such that that area is to the left of the point in that distribution.

$$a = \text{CHISQ.INV}(\text{area to the left of } a, d.f.) = \text{CHISQ.INV}(P(\chi^2 < a), d.f.)$$

Text Exercise 6.5.1

Within the context of constructing confidence intervals for variances from populations that are normally distributed. Determine the two critical values for the given confidence level and sample size by roughly sketching the χ^2-distribution and using technology.

1. $\text{CL} = 0.9$ and $n = 8$

Answer

We begin by sketching a χ^2-distribution with 7 degrees of freedom. It is a skewed right distribution starting at $0, 0$ with its peak at 5 (in general, at $(n-3)$). We then drawn our three regions and label them with the appropriate areas.

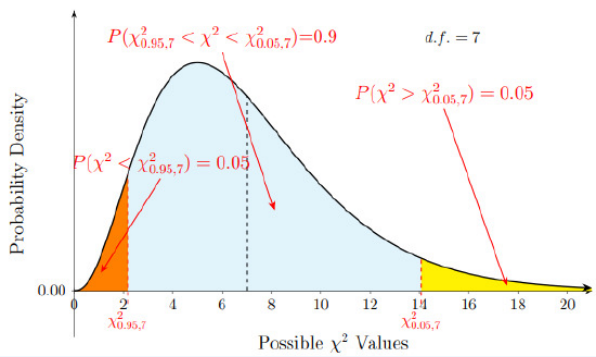


Figure 6.5.2 χ^2-distribution

$$\chi^2_{0.95,7} = \text{CHISQ.INV}(0.95, 7) \approx 2.6174$$

$$\chi^2_{0.05,7} = \text{CHISQ.INV}(0.05, 7) \approx 14.0671$$

2. $\text{CL} = 0.95$ and $n = 24$

Answer

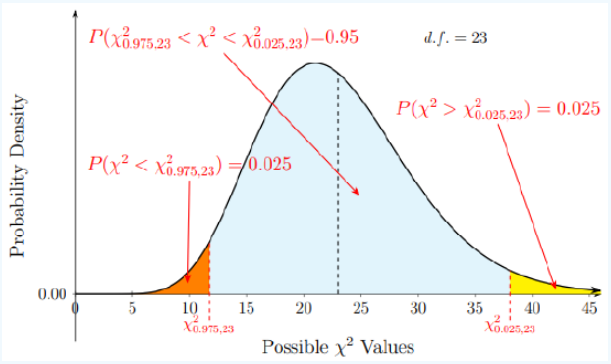


Figure 6.5.3 χ^2-distribution

$$\chi^2_{0.975,23} = \text{CHISQ.INV}(0.975, 23) \approx 11.6886$$

$$\chi^2_{0.025,23} = \text{CHISQ.INV}(0.025, 23) \approx 38.0756$$

3. $\text{CL} = 0.99$ and $n = 17$

Answer

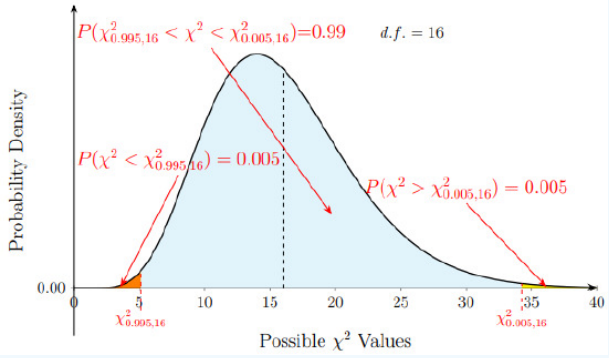


Figure 6.5.4 χ^2-distribution

$$\chi^2_{0.995,16} = \text{CHISQ.INV}(0.995, 16) \approx 5.1422$$

$$\chi^2_{0.005,16} = \text{CHISQ.INV}(0.005, 16) \approx 34.2672$$

Text Exercise 6.5.2

Let us return to the topic of the heights of adult males. In the previous section, we used a random sample of 10 adult males to construct a confidence interval for the population mean of adult males. Use the same sample data, provided again below, to construct a 90% confidence interval for the population variance.

$$64, 66, 67.5, 68, 69, 69.75, 71.25, 72, 73, 73.25$$

Answer

In order to construct a confidence interval for variances using the methods developed above, we need the parent population to be normally distributed and to use a randomly selected sample. The heights of adult males are normally distributed so we may construct our confidence interval.

To construct the confidence interval, we need the sample standard variance, sample size, and the two critical values. The sample variance of our particular sample is approximately 9.3924square inches. We sampled 10 adult males. To calculate the critical values, we must find α. Since $\text{CL} = 0.90, \alpha = 0.1$. Our critical values come from the χ^2-distribution with 9 degrees of freedom. We sketch the distribution to help us compute the critical values.

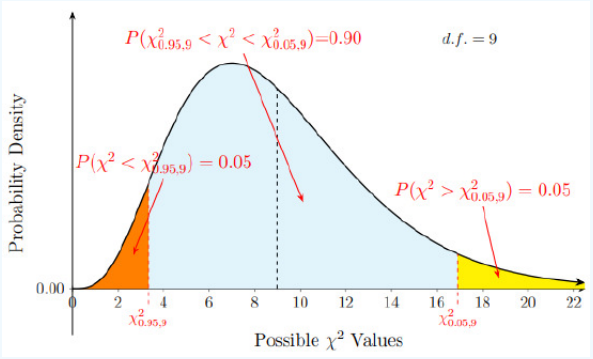


Figure 6.5.5 χ^2-distribution

$$\chi^2_{0.95,9} = \text{CHISQ.INV}(0.95, 9) \approx 3.3251$$

$$\chi^2_{0.05,9} = \text{CHISQ.INV}(0.05, 9) \approx 16.919$$

$$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{\text{CL}+\frac{\alpha}{2},n-1}} \right) \approx \left(\frac{9 \cdot 9.3924}{16.919}, \frac{9 \cdot 9.3924}{3.3251} \right) \approx (4.9962, 25.4221)$$

At the 90% confidence level, the population variance σ^2 of adult male heights is between 4.9962 square inches and 25.4221 square inches.

Constructing Confidence Intervals for Standard Deviations

We are often interested in the standard deviation of a population. Most of the theoretical work for sampling distributions and confidence intervals occurs within the realm of variance because the sample variance is an unbiased estimator of the population variance; while, the sample standard deviation is not an unbiased estimator of population standard deviation. But, we can use a confidence interval for variances to speak about standard deviations rather simply because the standard deviation is the square root of the variance, and $\sqrt{x}$ is an increasing function meaning that it preserves order. Let us return to the context of our last text exercise about adult male height. We are 90% confident that the population variance is between 4.9962 and 25.4221 square inches and can, therefore, be 90% confident that the population standard deviation is between $\sqrt{4.9962} \approx 2.2352$ and $\sqrt{25.4221} \approx 5.0420$ inches. This range of values is right where we might expect the population standard deviation of adult male heights given that the population standard deviation of adult female heights is about 2.5 inches.

The section concludes with a general formulation. If a parent population is normally distributed, we may, by randomly selecting a sample of size n from the population and setting a confidence level CL, construct a confidence interval for standard deviations of the form

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{\mathrm{CL}+\frac{\alpha}{2},n-1}}} \right)$$

where the critical values come from the χ^2-distribution with $n-1$ degrees of freedom.

CHAPTER OVERVIEW

7: Hypothesis Testing

7.1: Introduction to Hypothesis Testing

Learning Objectives

- Introduce the idea of hypothesis testing
- Define null and alternative hypotheses
- Develop the logic of identifying null and alternative hypotheses
- Define the p-value
- Explain the two possible conclusions for a hypothesis test
- Introduce the α value
- Introduce type I and type II errors
- Differentiate between statistically significant and practically significant results
- Introduce one-tailed and two-tailed tests

Review and Preview

In the last chapter, we finally achieved a goal of inferential statistics: to use facts about sample data to speak confidently about the facts of the population from which it was drawn. Our method of confidence intervals provides an interval estimate of the population parameter at a certain success rate called the confidence level. When we do not know much about the population, we can utilize random sampling to build confidence intervals to learn about populations from scratch. At other times, we have claims about a certain population that we hope to test. This quest falls within the realm of inferential statistics and is the subject of this chapter.

Consider the legendary secret agent James Bond 007, the central character of a series of books and action movies, who is capable of great feats of heroism and has a penchant for martinis: shaken not stirred, never stirred. With such a strong preference, one would expect that James Bond could actually tell the difference between shaken and stirred martinis simply by taste and know which was which. Such an ability may seem unlikely. It would be natural to desire some evidence to back the claim. Taste-testing martinis could serve as a method for collecting such evidence. Suppose we settled on conducting 16 taste tests where each martini was either stirred or shaken randomly. Upon tasting each martini, James Bond claimed the martini was either shaken or stirred. After tasting all 16 martinis, suppose James Bond correctly identified 13 of the martinis. Is this sufficient evidence to back the claim that he can distinguish between shaken and stirred martinis simply by taste? If instead he got all 16, would that prove that the claim is true?

Neither result proves the claim; we cannot construct proofs of such claims. He may have been merely guessing but had great luck. Is luck a plausible explanation? If we assume that he is merely guessing, what is the probability that he actually designates 13 of the 16 martinis correctly? What is the probability that he correctly assigns at least 13 of the 16 martinis correctly? This latter question covers the case of what actually happened and anything more extreme happening and will be a frequent concern throughout this chapter. We can understand the taste testing and probability questions with binomial random variables. We have 16 trials with a success defined as correctly classifying the martini as shaken or stirred. If we assume James Bond is guessing, then the probability of success for a single tasting is $p = \frac{1}{2}$. If X is the random variable counting the number of successes in 16 trials, we are interested in computing $P(X \geq 13)$. The reader is encouraged to verify that $P(X \geq 13) \approx 0.0106$. With such a small probability, we would say that someone randomly guessing at least 13 of the 16 martinis correctly is quite unusual. We have two possible situations: James Bond was guessing and something quite rare occurred by chance, or James Bond actually has the ability to distinguish shaken and stirred martinis by taste. Given the evidence, the claim that James Bond was merely guessing seems dubious; the claim has not been proven false, but there is considerable doubt about its validity. We, therefore, say that there is considerable evidence that James Bond can distinguish between martinis shaken and stirred simply by taste.

We call the process described above as hypothesis testing. There is a claim or hypothesis about reality that needs to be tested (that James Bond can distinguish between shaken and stirred martinis simply by taste). A competing hypothesis is identified (that James Bond was merely guessing). Under the assumption that the competing hypothesis is true (James Bond is guessing), the probability that what happened or even something more extreme will happen is computed $P(X \geq 13) \approx 0.0106$. If this is a rare event, it casts doubt on the validity of the assumption that the competing hypothesis is true; thus, credence is lended to the original hypothesis (that James Bond can indeed distinguish between shaken and stirred martinis simply by taste). At no point in this process, just as in

constructing confidence intervals, is certainty achieved. Rare things do indeed happen, but that does not mean we cannot have confidence in assessing evidence. We will now begin formalizing the process of hypothesis testing.

Hypotheses

As we interact with the world around us, we begin to notice pattens in our observations and start to form hypotheses about the world based on these patterns. Before we act as if these hypotheses are true, we want to secure reasonable and sufficient evidence in support of them. So, how do we collect evidence in support of a hypothesis that arose from the claims or observations of ourselves or others? Recall that, in our example with James Bond, we identified a competing hypothesis, a hypothesis that was opposite of the one he claimed. James Bond claimed to be able to distinguish between shaken and stirred martinis by taste. The opposite claim was that he could not distinguish by taste and, therefore, was guessing. We call these two competing hypotheses the alternative hypothesis H_1 and the null hypothesis H_0. The **null hypothesis** is more practically or reasonably assumed to be true. While the **alternative hypothesis** is generally the novel or claimed hypothesis.

Within the context of the James Bond scenario, our initial disposition was that having such a refined taste palette would be abnormal. It seemed more reasonable to assume, at least initially, that one would simply be left guessing about how the martinis were mixed. Thus, the null hypothesis was that James Bond could not distinguish simply by taste and, therefore, was guessing. The alternative hypothesis was that he could distinguish simply by taste.

Consider another example: preparing to enter the shower. An important concern is the temperature of the water. Either the water is amenable to a pleasant shower or it is not. We have two hypotheses. Which of the two hypotheses do we presume to be true as we prepare to enter the shower? Not many of us turn the water on and immediately hop in the shower. Instead, we wait for it to warm up, waiting to see steam rising or periodically running a hand through the water to test the temperature. These actions speak of an initial assumption that one of the hypotheses is true. We operate with the assumption that the water is unsatisfactory until we have evidence to the contrary. Our actions reveal that the null hypothesis is that the water is not amenable to showering.

$$H_0 : \text{Water is not at a temperature suitable to showering}$$
$$H_1 : \text{Water is at a temperature suitable to showering}$$

These designations could also be thought of in terms of the potential implications of acting as if one of the hypotheses is true when indeed it is not. If we assume that the water is ready but that is not the case. What might happen? Since we assume that the water is good to go, we will hop in the shower right away. Once we are in the shower, we quickly find out that it is either too hot or too cold and immediately feel something moderately unpleasant to possibly painful. If, on the other hand, we assume that the water is not yet ready despite the water actually being perfectly suitable to us, the price is that we wait an extra minute before hopping in the shower. Which of these situations would we prefer if we were wrong with our initial assumption? We would prefer just waiting around for an extra minute over something that could potentially scald us. The hypothesis with the less drastic cost in acting as if it were true when it was false is the null hypothesis.

? Text Exercise 7.1.1

Within each context, determine the competing hypotheses and identify them as either the null hypothesis or the alternative hypothesis. Explain your reasoning.

1. In a court of law in the United States, a prosecutor (a lawyer) argues that the defendant (a citizen) is guilty of some crime. The defendant is usually represented by a criminal defense lawyer who argues against the prosecutor (that his client is not guilty). A judge and possibly a jury follow the arguments in order to draw a final conclusion called a verdict.

Answer

The two competing hypotheses are related to the defendant's innocence regarding the crime charged against him. One hypothesis is that the defendant is innocent of the charge. The other is that the defendant is guilty of the charge. In the United States, our legal system is structured with the mantra "innocent until proven guilty," with sufficient evidence described as evidence beyond a reasonable doubt.

$$H_0 : \text{The defendant is innocent}$$
$$H_1 : \text{The defendant is guilty}$$

Thinking of the potential implications of acting as if one hypothesis were true when it is not can help us understand why our legal system was set up as it is. If we act as if the defendant is guilty despite the fact that he is innocent. We could send

an innocent person to pay a fine, spend time in prison, or even be executed. Our founders experienced tyranny and sought in many ways to protect the citizens from the government. The potential price from the defendant's perspective is quite steep. Sending an innocent man to jail for years or to death is severe. On the other hand, if we act as if the defendant is innocent despite the fact that he is guilty the potential price from the defendant's perspective is the cost of playing the system. It may be that the defendant feels remorse and a desire to change his ways or it may be that he got away with breaking the law. He may be more or less prone to be a repeat offender. The local society may deem acting with more caution around the defendant a prudent decision. From the perspective of our legal system, the cost of an innocent man being wrongly persecuted is worse than a guilty man walking free.

"It is better that ten guilty persons escape than that one innocent suffer." - William Blackstone

"It is better 100 guilty Persons should escape than that one innocent Person should suffer." -Benjamin Franklin

2. A researcher at Stine, a company that develops corn and soybean seeds, identifies a new breed of sweet corn in its breeding laboratory that he thinks will produce corn that is more tolerant to the dry conditions of northwestern Kansas than the breed that most northwestern Kansas corn farmers currently use.

Answer

The two competing hypotheses are related to the drought tolerance of the new variety of corn. One hypothesis is that the new variety of corn has better drought tolerance than the commonly used variety of corn. The other is that the new variety of corn does not have better drought tolerance than the commonly used variety of corn. This could be that it is equally tolerant or that it is less tolerant. The Stine company would benefit tremendously from developing a variety of corn suitable for drier conditions. If the commonly used seed is of a business competitor, Stine can expand its market. Even if it is the producer of the current common variety, the new seed will likely be able to be sold at a higher price and bring new attention to the company. Making the assumption that the new seed is better than the common seed could be disastrous if false. The company would likely lose many clients and possibly face a lawsuit for false advertising. As such, the reasonable hypothesis to assume initially until there is evidence to the contrary is that the new variety of seed is not better suited to the dry conditions of northwestern Kansas.

$$H_0 : \text{The new variety of sweet corn seed is, at best, the same as the commonly used seed}$$
$$H_1 : \text{The new variety of sweet corn seed is better than the commonly used seed}$$

Note: Pascal's Wager

The 17^{th} century philosopher and mathematician Blaise Pascal engaged in a similar line of reasoning before any rigorous development of hypothesis testing in an argument called Pascal's Wager which can be found in his book $Pensées$. Dr. Peter Kreeft, a renowned philosopher and professor at Boston College, gives an exposition of the argument (which can be found here) which we will formulate here using the framework of hypotheses that we have been developing.

Pascal lived in a time of great religious skepticism and attempted to formulate a line of reasoning that could reach a skeptic who lacked faith and did not believe that reason was sufficient to prove that God existed. There are two competing hypotheses at play: God exists and God does not exist. The skeptic knows that only one of the hypotheses is true but cannot establish intellectual certainty as to which to adopt. Pascal, just as is done within hypothesis testing, considers the potential ramifications of living as if the hypotheses were true when, indeed, they were false. Which hypothesis is it best to live by as if it were true? Pascal argues that one cannot abstain from the wager, for we are all already playing the game of life.

In considering the ramifications, Pascal, a Christian, considers the possibility of eternal happiness because everyone seeks maximal happiness. But in considering eternal happiness, he assumes that if God exists, there is paradise (heaven), thus equivocating the existence of God such that there is no paradise (heaven) with there being no God.

Pascal continues. If one lives as if God does not exist when He really does, then that one misses out on the possibility of eternal happiness. If one lives as if God does exist when He really does not, then that one has lost nothing since there was nothing to gain or lose at the moment of death. Which of the two of these potential costs is less drastic? He argues that the cost of eternal happiness is infinitely worse than losing nothing. Then we could say, within the context of competing hypotheses, that the null hypothesis, the hypothesis initially assumed and acted upon as true, would be that God exists.

$$H_0 : \text{God exists}$$
$$H_1 : \text{God does not exist}$$

Accepting the risk assessment of the two hypotheses is a critical part of his argument, and there are reasonable grounds to object and much to consider about the context, assumptions, nuances, and ramifications of the argument. An interested reader is encouraged to ponder the argument more thoroughly and read through Kreeft's exposition linked above.

Collecting Evidence and Making Decisions

In the example about martinis and James Bond, the null hypothesis was that James Bond simply guesses whether the martinis were shaken or stirred. We collected evidence to test his claim, the alternative hypothesis (that he could distinguish between shaken and stirred martinis simply by taste) by conducting an experiment where he taste tested 16 martinis. We then analyzed that evidence under the assumption that the null hypothesis was true by computing the probability that a person simply guessing would get at least 13 identifications correct when presented with 16 martinis where the mixing method was chosen randomly. This probability is called the ***p*-value.**

The p-value is not the probability that James Bond was simply guessing (i.e. it is not the probability that the null hypothesis is true). Rather, it is the probability that something at least as extreme as what was observed happens assuming that the null hypothesis is true. We can understand the p-value as the probability of an event given a hypothesis. It is often misunderstood as the probability of a hypothesis given an event.

James Bond correctly identified 13 of the 16 martinis. What would be at least as extreme in the context of the taste testing? It would be that he got $13, 14, 15,$ or even 16 martinis right. So in this context, the p-value is $P(X \geq 13)$. When the p-value is quite small, either something rare happened or there was a flaw in an assumption of our analysis, namely, that the null hypothesis is true. We cannot be certain which is the case from what we know, but if the p-value is sufficiently small, we generally consider it as significant evidence that the null hypothesis is false. When this is the case, we say that we **reject the null hypothesis in favor of the alternative hypothesis**. In the case of James Bond, the p-value was 0.0106, which is rather small, so we rejected the null hypothesis (that he was simply guessing) in favor of the alternative hypothesis (that he could indeed distinguish simply by taste). If the p-value is not sufficiently small, the event that occurred was not rare enough under the assumption that the null hypothesis is true to cast doubt on the truth of the null hypothesis. This does not prove that the null hypothesis is true; rather, we simply failed to show it was likely false, so we conclude that we **fail to reject the null hypothesis** when this happens. We emphasize that failing to reject is not the same thing as accepting. Hypothesis testing can only falsify, never verify, the null hypothesis, as throughout the procedure, the null hypothesis is assumed to be true.

Reports of statistical analyses outline the logical progression for the development of hypotheses, experimental design, results of the experiment, and the p-values in order to provide the readers with the full scope of the logic and evidence. This is done because there are different approaches to what is deemed a sufficiently small p-value and the decisions need to be based on more than just whether the p-value meets some given threshold. However, deciding on a threshold which considers the context and is determined independently from the evidence is a good start to measuring the weight of the evidence. Once a threshold is set, we describe the hypothesis test as having that level of significance, which is typically referred to as the **α value** of the hypothesis test. Commonly used α values are 0.05 and 0.01.

From our discussion thus far, we note that the null hypothesis plays a pivotal role in the process of hypothesis testing. The p-value comes from a probability calculation, assuming the null hypothesis is true. The conclusions that we draw from a hypothesis test come in two forms: reject the null hypothesis or fail to reject the null hypothesis. Indeed, we can loosely understand the process of hypothesis testing as the quest for finding evidence against the null hypothesis, so that, when significant evidence (evidence that produces a p-value less than the α value, the significance level) is found, we can reject the null hypothesis and favor the alternative hypothesis.

Type I and Type II Errors

Rare events are not impossible events; they do happen from time to time. As such, it is possible to conduct a hypothesis test, collecting evidence that meets our standard of significance, which leads us to reject a true null hypothesis. Perhaps, we are partially at fault for being too easily convinced that the evidence was strong. Perhaps something extraordinarily rare occurred, but we made an error either way. In rejecting a true null hypothesis, we made a statement about reality that does not match what really is happening. We call this error a **type I error**.

Recall that when deciding which of the competing hypotheses would be the null hypothesis, we were considering the potential ramifications of acting as if one of the hypotheses were true when it really was not. We could also consider the ramifications in the other direction. What are the costs of rejecting one of the hypotheses when it was true; in terms of our current discussion, what are the costs of committing a type I error if we adopt a particular hypothesis as our null hypothesis? Consider the showering example once more. If we reject that the water is suitable, when it is not, the cost is the more drastic of the two. That was the hypothesis we set as the null hypothesis. Thus, the hypothesis with a more severe cost in rejecting it when it is true is classified as the null hypothesis. We do this because we have a certain degree of control over the occurrence of type I errors; we set the thresholds regarding sufficient evidence.

We cannot completely avoid making a type I error, but we can manage the likelihood. As we have seen, evidence is collected, the p-value is computed using the evidence, and then if the p-value is less than the α value (p-value$< \alpha$), we assert that there is sufficient evidence to reject the assumption that the null hypothesis is true. The α value is the upper bound regarding which p-values accounted for sufficient evidence, and recall that the p-value is the probability that something at least as extreme as what happened happens given the assumption that the null hypothesis is true. So, we can understand α as the probability of making a type I error. A smaller α value means a smaller rate of committing type I error.

Making a type I error is not the only way in which we may fail to recognize the reality of the world around us. It may be the case that the null hypothesis is false and would be rejected, but we fail to do so because the evidence did not meet the level of significance desired. Recall that, when the evidence gathered is not sufficient, the result of the hypothesis test is that we fail to reject the null hypothesis. Failing to reject the null hypothesis is not the same as declaring the null hypothesis is true. We are not making a strict assertion regarding which hypothesis matches reality; we are simply saying that the evidence did not cast sufficient doubt on the truthfulness of the null hypothesis. Despite the fact that, in failing to reject a false null hypothesis, we are not asserting anything false about reality, we still call it an error in the fact that we have failed to be better aligned with reality. We call this error a **type II error**.

Once again, consider the showering example. We set up the hypotheses as follows.

$$H_0 : \text{Water is not at a temperature suitable to showering}$$
$$H_1 : \text{Water is at a temperature suitable to showering}$$

A type I error would mean falsely believing that the water is suitable. A type II error would mean falsely believing that the water is not suitable. Notice that if we switched which hypothesis was the null hypothesis, the error types would also switch. The fact that our framing yields a type I error which is more severe than the type II error indicates the hypotheses were formulated correctly.

Note: Hypotheses and Errors

Our original discussion on determining which hypothesis to set as the null hypothesis centered on the ramifications of acting as if one of the hypotheses were true when that was not the case. We set the null hypothesis as the hypothesis with the less drastic costs; the alternative hypothesis would thus have the more drastic costs. This process is similar to the consideration of making a type II error. In failing to reject a null hypothesis which is false, we do not assert that the null hypothesis is true, but our initial disposition towards the hypotheses remains the same. We can connect the two ideas to formalize our hypothesis-setting process; the alternative hypothesis is set by considering the ramifications of making a type II error if the particular hypothesis is set as the alternative hypothesis. The hypothesis with the more drastic cost in making a type II error is the alternative hypothesis. We have two mechanisms for deciding how to set the hypotheses in a hypothesis test (both produce the same results).

1. When considering which hypothesis to set as the null hypothesis, consider the costs of committing a type I error if the hypotheses were adopted as the null hypothesis. Set the hypothesis with the greater cost as the null hypothesis.
2. When considering which hypothesis to set as the alternative hypothesis, consider the costs of committing a type II error if the hypotheses were adopted as the alternative hypothesis. Set the hypothesis with the greater cost as the alternative hypothesis.

Both methodologies produce hypotheses such that a type I error is worse than a type II error. This is because we can precisely identify the probability of a type I error, but we cannot control the likelihood of a type II error. One can verify that one has correctly set up hypotheses by checking that a type I error is worse than a type II error.

Conducting Hypothesis Testing

In inferential statistics, we are primarily interested in claims about populations and assertions about the values of parameters. We could test the claims with certainty if we could compute the actual parameter value, but alas, that is not the case in reality due to literal or practical impossibilities. We use random sampling to collect evidence and then use our knowledge of sampling distributions to compute the p-value.

Consider a seed company with a new sweet corn variety that is thought to produce greater yields in northwestern Kansas than what is commonly planted now. We can measure such a claim by looking at the average yield of the two varieties; these are facts about all seeds of the two varieties, i.e., parameters. After seasons of planting and studying the common sweet corn variety, researchers and farmers have a pretty good idea about its average yield; let us suppose that it is equal to 70 bushels per acre. Since the researchers at the seed company think that they have discovered a better-yielding variety of sweet corn, we can assume there is a line of scientific reasoning or some test plots that have led to such a hypothesis, but we do not know the value of the average yield for the new variety; let us simply call it μ. Let us suppose, for pedagogical purposes, that the new sweet corn varieties from the company consistently have standard deviations of 3.6 bushels per acre; so, that we feel confident the new variety has a standard deviation of 3.6 bushels per acre too.

Just as in the case with the drought-resistant new variety of seed, we do not want to assume initially that the new seed is better than the commonly used seed. So, our null hypothesis is that the average yield of the new variety is, at best, the same as the average yield of the commonly used variety, and our alternative hypothesis is that the average yield of the new variety is better than the average yield of the commonly used variety. We can denote this symbolically as follows.

$$H_0 : \mu \leq 70 \text{ bushels per acre}$$
$$H_1 : \mu > 70 \text{ bushels per acre}$$

Let us suppose that the company has set a standard α value for these sorts of tests at the 0.01 level. With the two hypotheses and α value set, we now look to gather evidence by randomly sampling the yields of the new variety of sweet corn. Perhaps we randomly sample 50 acres across northwestern Kansas where the new variety was grown under comparable conditions to the normal farming practices. From this random sample, we compute the average yield of the sample to be 71 bushels per acre. The average yield of the sample is greater than the commonly grown average yield, but is it sufficient evidence to cast doubt on the truth of the null hypothesis?

How do we assess our evidence? We compute the p-value, the probability that something more extreme happens than what we observed given that the null hypothesis is true. We are interested in computing $P(\bar{x} > 71 \text{ bushels per acre})$. Since we randomly sampled using a sample size of 50, we expect the sampling distribution of sample means to be approximately normal with a mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. From the consistency of the company's seeds, we can feel confident that $\sigma_{\bar{x}} = \frac{3.6}{\sqrt{50}}$. But, from the null hypothesis, we do not know precisely what μ is supposed to be. All we know from the assumption that the null hypothesis is true, is that $\mu \leq 70$ bushels per acre. A common adage tell us to prepare for the worst and hope for the best. We similarly want to consider the case where sufficient evidence against the null hypothesis would be hardest to obtain. Indeed, this is when we assume the new breed produces just as well as the commonly used variety, when $\mu = 70$ bushels per acre. This happens to be the mu value that produces the largest p-value (think about why this is the case). We encourage the reader to verify that the p-value is approximately 0.0248.

A p-value of 0.0248 would constitute significant evidence at the $\alpha = 0.05$ level but would not constitute significant evidence at the $\alpha = 0.01$. This seed company thus would fail to reject the null hypothesis in this situation. That is not to say that the new variety is not better than the common seed. There has not been sufficient evidence to doubt that the new yields are, on average, comparable to the common seed. If the research team is confident in the genetics of their newly developed variety, it may be prudent to conduct another experiment with a larger sample (recall that the standard deviation of the sampling distribution shrinks as the sample size increases) because rare things do occur.

? Text Exercise 7.1.2

Suppose the researchers at the seed company were very confident in the scientific reasoning behind the genetic breakthrough with this most recently developed sweet corn variety, and they decided to conduct an experiment using 10 times as much data, meaning, 500 acres were to be randomly selected across northwestern Kansas to grow this new variety. After the harvest, the researchers found that the average yield from these 500 acres was 70.4 bushels per acre.

1. Using the same hypotheses as the example above, determine the p-value from this larger experiment. State and interpret the conclusion of the hypothesis test at the $\alpha = 0.01$ level.

Answer

Since the hypotheses from the first experiment have not changed, the p-value is the probability that a random sample of 500 acres produces an average yield at least as large as the average yield that was observed. So we are looking to compute $P(\bar{x} > 70.4 \text{ bushels per acre})$. The sampling distribution is approximately normal because the sample size is so large with a standard deviation of $\sigma_{\bar{x}} = \frac{3.6}{\sqrt{500}}$. We will again assume that $\mu = 70$ bushels per acre because that is the condition within the assumption that the null hypothesis is true that will produce the largest p-value. p-value $= 1 - \text{NORM.DIST}(70.4, 70, \frac{3.6}{\sqrt{500}}, 1) \approx 0.0065$.

Since $0.0065 < 0.01$,we say that at the significance level of 0.01 there is sufficient evidence to reject the null hypothesis that the average yield of the new variety of sweet corn yields, at best, the same as the commonly used corn seed. We conclude that the new variety of sweet corn produces, on average, a greater yield than the commonly used seed.

2. At the time of the study, sweet corn was being sold at $\$4.20$ per bushel on average. If a farmer with 80 acres designated for sweet corn was seeking advice about switching to the new breed of corn, what aspects of the study would be important to consider? What would your advice to the farmer be?

Answer

Since we rejected the null hypothesis at the α level of 0.01, there is statistical evidence to say that the average yield of the new variety of sweet corn is larger than 70 bushels per acre, the average yield of the commonly used variety. The hypothesis test itself did not determine by how much the average yield would increase; it only concluded that there was an increase. If the increase was large and resulted in larger profits despite having to pay more for the seed, the farmer may be inclined to make the switch. If the increase was not large, the farmer may be less profitable despite producing more because of the increased cost associated with the new seed. If we used the average yield from the sample to make such a comparison, we would expect the farmer to have an increased yield of 0.4 bushels per acre on average. We could then estimate the increased revenue to be $0.4 \cdot 80 \cdot 4.20 = 134.4$ dollars. Not knowing the actual prices of seed, we cannot estimate the profit, but it seems doubtful that the switch will lead to any increased profits worth noting.

More technical answers for the mathematically inclined.

Rather than just using the sample mean as a point estimate for the population mean of the new variety, we can construct a confidence interval and use the boundary points as upper and lower bounds to create an interval estimation of the increased revenue. Let us construct a 95% confidence interval; $70.0845, 70.7156$So, at the 95% confidence level, we expect the population mean to fall between $70.0845$and 70.7156bushels per acre. This leads us to expect to increase the yield by something between $0.0845$and 0.7156bushels per acre, meaning the expected increase of revenue would be between $\$28.38$and $\$240.42$which again does not seem that a switch would lead to any increased profits worth noting.

As we have just seen, there may be differences that are statistically significant that do not warrant changes in our lives. When this is the case, we say that the findings are statistically significant but not practically significant. Remember that we are trying to understand the world better so that we may better live in the world and interact with the world. The p-value does not measure the size of a difference, often referred to as the size of the effect; it measures the probability that something at least as extreme happens as what was observed to happen under the assumption that the null hypothesis is true. A small p-value does not indicate a large effect. Many people have fallen prey to misunderstanding the meaning and uses of hypothesis testing, especially regarding p-values. In fact, widespread misinterpretations of the p-value prompted the American Statistical Association published an article addressing the misuses of p-values in 2016 to help remedy the issue.

Types of Hypothesis Tests

Let us return to the James Bond example. What would we have concluded if, after tasting the 16 martinis, James Bond was only correct on 3 of them? If he were truly guessing, we would expect that he would be correct about half of the time, but that is not what happened. Such a low score could indicate that James Bond could taste a difference between shaken and stirred martinis, but

he has confused the tastes. He consistently labels shaken as stirred and stirred as shaken. This would indicate that he can taste a difference, but his preference for shaken is mistaken!

If we were simply interested in whether James Bond's taste buds could tell the difference between shaken and stirred martinis, and not that he could distinguish between them due to his strong preference for shaken martinis, evidence against the null hypothesis that he was guessing would come in the form of either really low numbers of successes or really high numbers of successes. We are looking for evidence in two different directions. In this case, looking at the probability of something at least as extreme happening as what happened takes on a more complicated meaning.

With 13 out of 16 taste tests being successful, we easily understand more extreme as at least 13 martinis being correctly identified. But what would be as extreme in the other direction in which we are looking for evidence? That would be 3 out of 16 taste tests being successful. And in this direction, at least as extreme would lead us to consider at most 3 martinis being correctly identified. So, we are looking at the two tails of the binomial distribution to compute the p-value. We call such a test a **two-tailed test**. Thus the p-value is $P(X \leq 3 \text{ or } X \geq 13) \approx 0.0212$. This, again, qualifies as sufficient evidence at the α value of 0.05.

In our previous examples, we have looked for evidence only in one direction. The p-value came from only one of the tails of the probability distribution; we call such tests **one-tailed tests**. Two-tailed tests are most common in scientific research because finding any difference is generally notable, interesting, and could lead to further development. In the case of medicine and business, one-tailed tests arise with greater frequency because there are cases where there is no need to distinguish between no effect and an effect in an unpredicted or undesired direction. For example, with the varieties of corn, the seed company is interested in increasing the yield. If a new breed has the same yield or a worse yield, the new breed does not warrant further consideration.

Note: Two-Tailed Tests and Their Conclusions

Consider the example with James Bond without considering his preference for shaken martinis. Does he possess the taste buds necessary to note the difference between the mixing method? We identified this as a two-tailed test and could write the hypotheses as follows.

$$H_0 : p = 0.5$$
$$H_1 : p \neq 0.5$$

The evidence from the taste tests remains the same; he correctly identified 13 out of the 16 martinis which produces a p-value of $P(X \leq 3 \text{ or } X \geq 13) \approx 0.0212$. At the $\alpha = 0.05$ level, there is sufficient evidence to reject the null hypothesis that he is merely guessing. We could state the conclusion as two sided: he can distinguish (perhaps incorrectly) between shaken and stirred tastes; or, given that he correctly identified most of the drinks, we could state a stronger conclusion: he can correctly identify if a drink is shaken or stirred. We will argue that the latter conclusion is not the proper inference given how the alternative hypothesis was formulated.

The statistician Kaiser published a paper in 1960 arguing that we can make the claim that James Bond can correctly distinguish between shaken and stirred martinis simply by taste. Some textbooks argue this is permissible; others argue that it is not. This alternative hypothesis looks like $H_1 : p > 0.5$ rather than $H_1 : p \neq 0.5$ as originally formulated. The form of the alternative hypothesis changed to match the direction of the sample statistic. We will not adopt this practice. Since we are operating within the realm of formal hypothesis testing, we will maintain the form of the original alternative hypothesis and conclude that $p \neq 0.5$.

To understand our position, we further explain the process and purpose of hypothesis testing. Hypothesis testing is meant to test hypotheses formulated from previous observations, previous experimentation, or working theories. Once the hypotheses have been set, new experimentation is implemented, and the results are used as evidence in the hypothesis testing. The formulation of the hypotheses happens before the experimentation and is independent of the new data. That is not to say that the data from the new experiments cannot be studied to formulate further hypotheses or nuance the current hypotheses. This is to say that these new hypotheses are not to be tested using the same data precisely because the data led to them. That would be circular reasoning. Strictly speaking, new experiments need to be conducted to test the new hypotheses. This includes modifying the alternative hypothesis of a two-tailed test into a one-tailed alternative hypothesis.

This consideration is closely related to a problem plaguing much of modern academia, where publishing statistically significant results often takes precedence over honest intellectual inquiry, resulting in the temptation to conduct unplanned analyses of experimental data in search of a statistically significant result. Acting on this temptation is called p-hacking and is not an ethical research practice. When tests are done on patterns already seen in the data, the tests become meaningless. The proper

path forward is to study the experimental data, formulate new hypotheses, and then test those hypotheses via new experimentation.

Let us consider the case of James Bond once more. Suppose that James Bond could tell the difference by taste but did not have the tastes correctly aligned with the methods. But, somehow, he still managed to get 13 of the 16 martinis labeled correctly. In this case, it is even less probable that he performed as observed than when we just assumed he was guessing. The p-value using both tails was already small enough to warrant rejection at the particular significance level; so, this is evidence that we would expect the parameter p to fall on the same side of the value in the null hypothesis 0.5 as the sample statistic. We agree that it is very tempting to conclude $p > 0.5$. Perhaps in the messiness of practical application, action may to be taken based on that conclusion, but in the rigors of hypothesis testing, another experiment is warranted to test the claim on untested data that is randomly sampled.

There is contention in the textbooks; perhaps, there is less across the field of active statisticians. Interested readers are encouraged to deepen their understanding of both sides of the argument. For an argument to the contrary of our position, see Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review*, *67*, 160-167.

7.2: Claims on Population Means

Learning Objectives

- Test claims on population means both when σ is known and when σ is unknown
- Generalize the three forms of hypothesis tests
- Introduce, motivate, and utilize test statistics in computing p-values

Section 7.2 Excel File: (contains all of the data sets for this section)

Means and Hypothesis Testing

Now that we have been introduced to the general logic of hypothesis testing, we will begin to address the particulars found within hypothesis testing based on the parameters of interest. We begin with testing hypotheses about the population mean. Just like our considerations of confidence intervals for means, we will have two cases to consider based on whether the population standard deviation is known or unknown. The latter case is the more frequently occurring case as we discussed in the chapter on confidence intervals, but we again begin with the case that the population standard deviation is known because of pedagogical considerations.

Recall the general process of hypothesis testing. A claim is made that warrants testing. This hypothesis may be derived from simple observation, past experimental data, a person, or an institution, and it tends to be the alternative hypothesis. A competing hypothesis, which tends to be the null hypothesis, is constructed. The statement of the alternative hypothesis determines the type of test to be conducted (either a one-tailed test or a two-tailed test). At this point, researchers generally settle on the level of the test (the probability of a type I error given that the null hypothesis is true). The next step in the process is designing the experiment to ensure that the test can actually be conducted. The calculation of the p-value depends on the sampling distribution of the sample statistics used to estimate the population parameter, assuming the null hypothesis is true. We need to ensure certain conditions are met. When testing claims on population means, we utilize the sampling distribution of sample means, which is normal when the underlying population is normal, and approximately normal for the most common distributions (recall our previous discussion on sampling distributions) when the sample size is larger than 30. We also need to ensure that our sample is randomly selected. Once we have designed the experiment, it must be conducted and analyzed. We then compute the p-value, the probability that at least as extreme as what was observed happens. If the p-value $< \alpha$ value, then we conclude that there is sufficient evidence to reject the null hypothesis. If the p-value $\geq \alpha$ value, then we conclude that there is not sufficient evidence to reject the null and hence fail to reject the null hypothesis. When sharing the results of the hypothesis test, include the p-value so others can also assess at the desired level of significance. With this succinct review, let us enter into testing claims on population means when σ is known.

Claims on Population Means (σ known)

Text Exercise 7.2.1

A bottling company is responsible for bottling 2 liter bottles of Dr. Pepper. The company policy regarding quality assurance is required to randomly sample 100 bottles each week to assess how well the bottles are being filled. The company assesses the test at a significance level of 0.01. If the company that built the bottling equipment guarantees that the machinery operates with a standard deviation of 0.1 liters and the last sample of 100 randomly chosen bottles had a sample mean of 1.98 liters, determine the hypotheses, make a conclusion regarding the test, and interpret the meaning within the context of the problem.

Answer

Since we are dealing with the amount of Dr. Pepper filled in the 2 liter bottles, we are dealing with population means. Each bottle has a certain amount of Dr. Pepper. Each bottle is supposed to have 2 liters of soda. So the population mean should be 2 liters. One hypothesis would be $\mu = 2$ liters. The bottling company wants to make sure that it is not overfilling (it does not want to shrink its profit margin) nor underfilling (it does not want to upset its customers and reface any false advertising lawsuits). We conclude that the other hypothesis would be $\mu \neq 2$ liters.

We now need to determine which hypothesis is to be the null hypothesis. If the company assumes that the machinery is not filling properly from the outset, it would automatically recalibrate the machinery every time; the collecting of evidence would be unnecessary. If the company acts as if the machines are not working properly when they really are, the company would unnecessarily waste production time. On the other hand, if the company acts as if the machines are filling properly

when they really are not, the company would produce bottles without the proper amount in them. If it was enough that customers would notice, the employees would likely notice before shipping them out. We, therefore, set the hypotheses as follows.

$$H_0 : \mu = 2 \text{ liters}$$
$$H_1 : \mu \neq 2 \text{ liters}$$

Since we had a random sample of 100 bottles, we have that the sampling distribution of sample means is approximately normal. Since we assume that the null hypothesis is true for the computation of the p-value, we have that the mean of the sampling distribution of sample means is $\mu_{\bar{x}} = 2$ liters. We also have that the standard deviation of the sampling distribution of sample means is $\sigma_{\bar{x}} = \frac{0.1}{\sqrt{100}} = \frac{0.1}{10} = 0.01$. Since our alternative hypothesis is that the mean is not equal to 2 liters, we are looking for evidence in two directions; we have a two-tailed test. Recall that the sample mean was 1.98 liters, which is less than 2 liters. We need to find the value that would be just as extreme except in the opposite direction. Since 1.98 liters is 0.02 liters below the hypothesized value, the other value we are looking for is 0.02 liters above the hypothesized value, namely 2.02liters. To find the p-value, we compute the area in the left tail ending at 1.98liters and the right tail starting at 2.02liters. See the figure below for a visual.

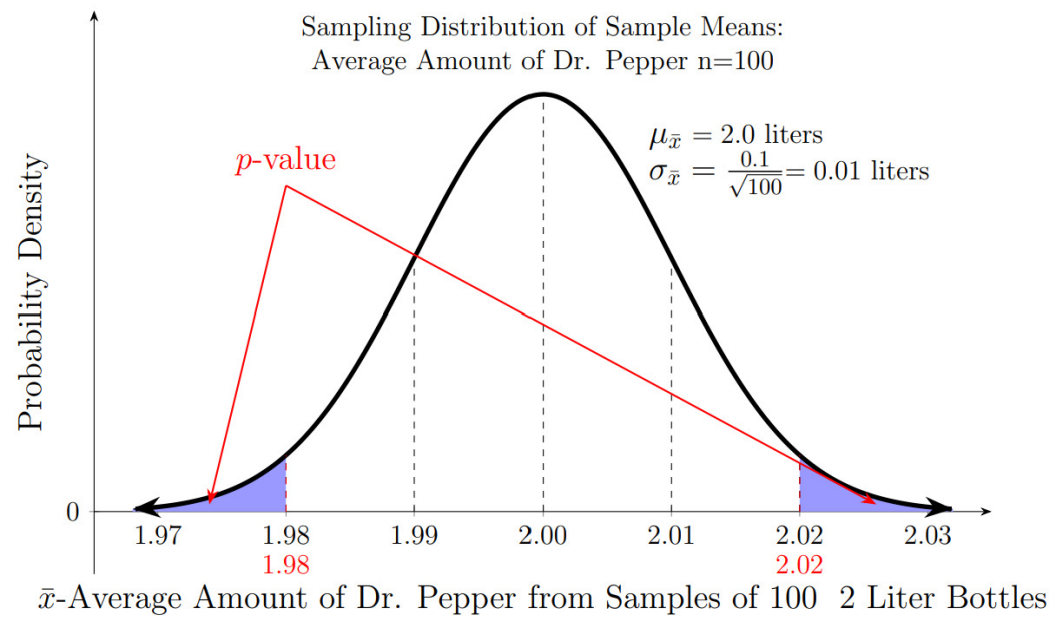


Figure 7.2.1: Sampling distribution of sample means

We compute the p-value using technology. Note that since the sampling distribution of sample means is approximately normal, the area in the two tails is equal due to symmetry. This eases the calculation. For this first exercise, we compute it both ways.

$$\begin{aligned} p\text{-value} &= \text{NORM.DIST}(1.98, 2, 0.01, 1) + (1 - \text{NORM.DIST}(2.02, 2, 0.01, 1)) \\ &\approx 0.02275 + (1 - 0.97725) \\ &\approx 0.0455 \\ p\text{-value} &= 2 \cdot \text{NORM.DIST}(1.98, 2, 0.01, 1) \\ &\approx 2 \cdot 0.02275 \\ &\approx 0.0455 \end{aligned}$$

Since 0.0455is not less than 0.01,we do not have sufficient evidence to reject the null hypothesis that the machines are filling the 2 liter bottles properly. The machines, therefore, pass the weekly test for quality assurance.

Text Exercise 7.2.2

In 2024, researchers at the University of Maryland edited the genes of poplar trees to reduce the amount of lignin naturally present in the tree. This is desirable because the process of strengthening wood involves heat and compression, and the more lignin present in the tree, the harder the tree is to compress. The amount of lignin present in a poplar tree is often reported as a percent of the tree's dry weight. The average poplar tree has 27% of its dry weight due to the weight of lignin.

Suppose that independent researchers want to verify the claim. They request to randomly sample 45 of the many thousands of genetically altered poplar trees growing across the various stations the University of Maryland researchers utilize. For the sake of open and honest scientific research, their requests are granted. If it is known that the standard deviation of the percent of weight of lignin in all poplars, including this genetically altered poplar, is 4.2% and the sample mean was 25.3%, conduct the hypothesis test at the 0.005 level of significance.

Answer

The measure of interest for each tree is the percent of dry weight that is due to the presence of lignin. Despite the fact that this measurement returns a percent, we are not considering proportions as our parameter of interest. We are interested in the population mean of the percent weight due to lignin measurements in the genetically altered poplar trees. It is known that the typical poplar tree has 27% of its dry weight due to the weight of lignin, and the researchers at the University of Maryland think that they have reduced the amount of lignin naturally present in the genetically altered poplar tree. We naturally obtain the hypothesis that $\mu < 27$ percent. From here, we identify the opposing hypothesis that $\mu \geq 27$ percent. Note that this is a one-tailed test. We conduct a one-tailed test because regardless of whether the genetically altered poplar trees have the same amount of lignin or more lignin than the regular poplar trees, interest in these genetically altered trees would fade. There is no need to distinguish between no change and a change for the worse.

To select which hypothesis is to be considered the null hypothesis, we note that we are conducting the experiment to test that the claims of the University of Maryland researchers are true; so, we do not want to assume their conclusion from the beginning. We will set the null hypothesis to say that the genetically altered poplars have at least as much lignin present as a percent of their dry weights as regular poplar trees.

$$H_0 : \mu \geq 27 \text{ percent}$$
$$H_1 : \mu < 27 \text{ percent}$$

We now begin to look at the sampling distribution of sample means given the size of the random sample and the assumption that the null hypothesis is true. Since the sample size is 45, we expect the sampling distribution of sample means to be approximately normal. Since there are many possible population means under the assumption that the null hypothesis is true, we conduct our study with the value of μ that will produce the largest p-value. This occurs when $\mu = 27$ percent.

We have $\mu_{\bar{x}} = 27$ percent and $\sigma_{\bar{x}} = \frac{4.2}{\sqrt{45}} \approx 0.6261$percent. We need to determine what would be considered at least as extreme as the evidence from the sample which produced a sample mean of 25.3percent. In this case, the more extreme would be smaller and smaller percentages. So we are looking for the area in the tail on the left side of the sampling distribution that ends at 25.3 percent. Notice how the direction of the tail matches the direction of the inequality in the alternative hypothesis; we call this a left-tailed test. See the figure below for a visualization.

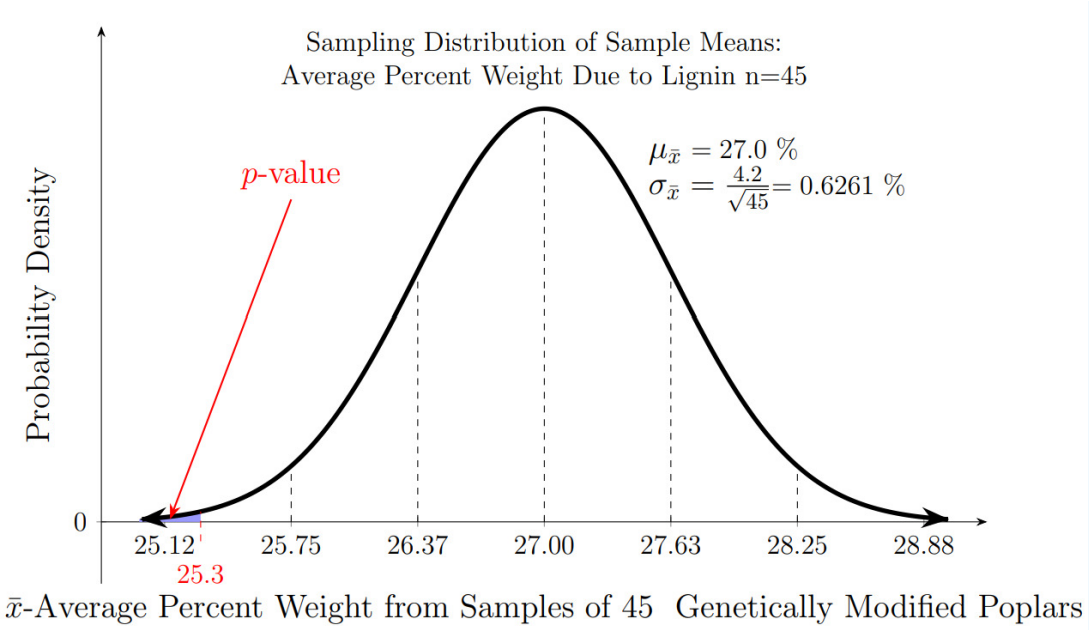


Figure 7.2.2 Sampling distribution of sample means

We thus compute the p-value using technology.

$$p\text{-value} = \text{NORM.DIST}(25.3, 27, \frac{4.2}{\sqrt{45}}, 1) \approx 0.0033$$

Since $0.0033 < 0.005$, there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis that the genetically altered popular trees have less lignin naturally present. There is sufficient evidence in support of the claims of the researchers at the University of Maryland.

Note: Hypotheses and p-values

At this stage, we have conducted both types of one-tailed tests: right-tailed (recall the last section with corn yields) and left-tailed (this section with lignin). When we operated under the assumption of the truth of the null hypothesis, we had to decide which value of μ to use in computing the p-value. We chose the value that would be the hardest to find sufficient evidence (the value that would produce the largest p-value). This turned out to be when μ equaled the common or accepted value from the problem (the 70 bushels per acre of the commonly used corn and the 27% of the regular poplar trees). This will always be the case and contributes to some textbooks making the pedagogical choice to say that the null hypothesis is always of the form: the parameter equals the standard value. We do not adopt this simplification for the form of the null hypothesis, but we do emphasize that when the parameter equals the standard value, the p-value produced will be the largest, which is essential for our analyses.

We also want to emphasize that there are three tests to consider: left-tailed tests, right-tailed tests, and two-tailed tests. These three tests correspond to the three possible forms of hypotheses. Consider looking at the alternative hypotheses. In this last text exercise about poplar trees, the alternative hypothesis was $\mu < 27$ percent, and we computed the p-value by calculating the area in the left tail of the distribution. In the text exercise about Dr. Pepper, the alternative hypothesis was $\mu \neq 2$ liters, and we considered the area in both tails. And finally, in the exercise about corn yields, the alternative hypothesis was $\mu > 70$ bushels per acre, and we looked only at the area in the right tail. Notice that the direction of the tail matches the direction of the inequality sign in the alternative hypothesis. All of this is provided in the figure below within the context of means but applies to any parameter (μ_0 indicates the common or accepted value of the population mean).

left-tailed test	two-tailed test	right-tailed test
$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$
use $\mu = \mu_0$ to compute p-value	use $\mu = \mu_0$ to compute p-value	use $\mu = \mu_0$ to compute p-value

Test Statistics

Recall that we can transform any normal distribution into the standard normal distribution and that this transformation preserves area. An implication of these facts is that the computation of p-values can be done within the context of the standard normal distribution once we transform the particular sample mean. We call this transformed value of the calculated sample statistic the **test statistic**. As discussed, different transformations can change sampling distributions of particular sample statistics to particular common distributions. For now, understand that the basic idea of the test statistic is that it is a value that represents the value of the sample statistic computed from the actual sample collected which facilitates computing the p-value.

Recall that the z-score transformation sends any normal distribution with a mean μ and a standard deviation σ to the standard normal distribution given by the formula below.

$$z = \frac{x - \mu}{\sigma}$$

? Text Exercise 7.2.3

Repeat the hypothesis tests from Text Exercises 7.2.1 and 7.2.2 using test statistics to compute the p-values. Verify that the same p-values are computed which in turn yield the same conclusions as before.

1. The first text exercise considered filling 2 liter bottles of Dr. Pepper. A sample of 100 2 liters was randomly chosen which produced a sample mean of 1.98 liters. The population standard deviation was 0.1 liters. The hypothesis test was to be conducted on the hypotheses below at the $\alpha = 0.01$ level of significance.

$$H_0 : \mu = 2 \text{ liters}$$
$$H_1 : \mu \neq 2 \text{ liters}$$

Answer

All conditions to conduct a hypothesis test are met; for details, review Text Exercise 7.2.1. We need to compute our test statistic. We assume that the null hypothesis is true and, therefore, know that the sampling distribution is approximately normal with $\mu_{\bar{x}} = 2$ liters and $\sigma_{\bar{x}} = \frac{0.1}{\sqrt{100}} = 0.01$ liters. Since we are transforming the sampling distribution of sample means with σ known into the standard normal distribution our z-transformation takes on the form below.

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

We can insert the values from our particular context to arrive at the following.

$$z = \frac{1.98 - 2}{\frac{0.1}{\sqrt{100}}} = \frac{-0.02}{0.01} = -2$$

Understand the -2 value to mean that under the assumption of the truth of the null hypothesis the evidence that we collected from the sample mean is 2 standard deviations below the hypothesized population mean.

The alternative hypothesis contains the $\neq$ sign implying a two-tailed test. We need to consider what is equally extreme in the opposite direction. Since the standard normal distribution is centered at 0, all we have to do is take the value equal in magnitude and opposite in sign, 2. Two values are equally extreme if they are the same number of standard deviations away from the population mean. We have the following visualization for computing the p-value.

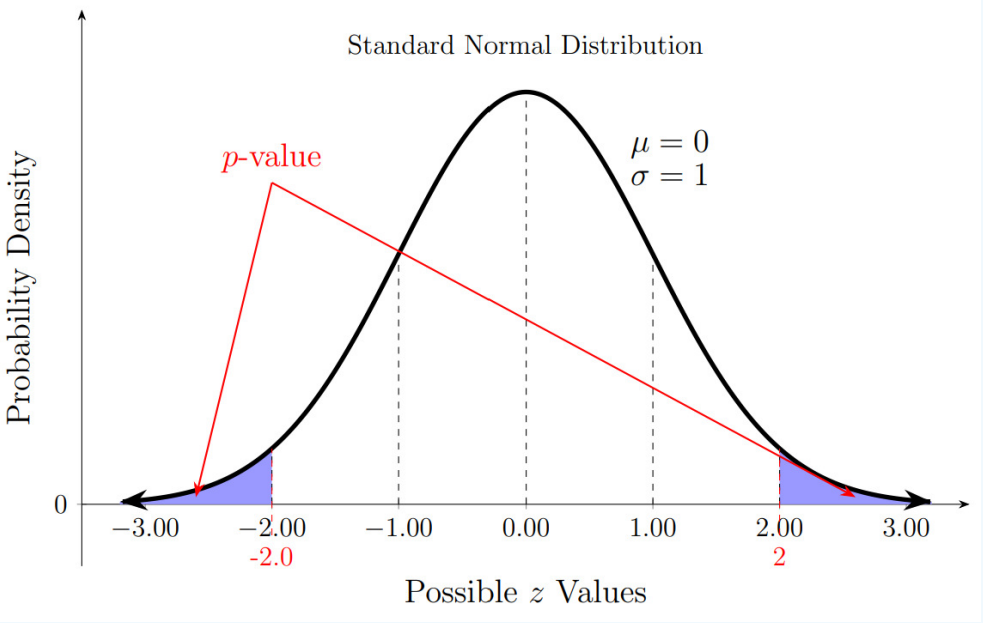


Figure 7.2.3 Standard normal distribution

$$\begin{aligned} p\text{-value} &= \text{NORM.S.DIST}(-2,1) + (1 - \text{NORM.S.DIST}(2,1)) \\ &\approx 0.02275 + (1 - 0.97725) \\ &\approx 0.0455 \\ p\text{-value} &= 2 \cdot \text{NORM.S.DIST}(-2,1) \\ &\approx 2 \cdot 0.02275 \\ &\approx 0.0455 \end{aligned}$$

We arrive at the same p-value as before and since $0.0455$is not less than 0.01,we fail to reject the null hypothesis.

2. The second text exercise considered the percent weight of poplar trees due to lignin. A sample of 45 genetically altered poplar trees was randomly chosen, producing a sample mean of 25.3 percent. The population standard deviation was 4.2 percent. The hypothesis test was conducted on the hypotheses below at the $\alpha = 0.005$ significance level.

$$\begin{aligned} H_0 &: \mu \geq 27 \text{ percent} \\ H_1 &: \mu < 27 \text{ percent} \end{aligned}$$

Answer

All of the conditions to conduct a hypothesis test are met; for details, review Text Exercise $7.2.2$.We need to compute our test statistic. We assume that the null hypothesis is true and, therefore, know that the sampling distribution is approximately normal with $\mu_{\bar{x}} = 27$ percent and $\sigma_{\bar{x}} = \frac{4.2}{\sqrt{45}} \approx 0.6261$percent. We can compute our test statistic.

$$z = \frac{25.3 - 27}{\frac{4.2}{\sqrt{45}}} \approx \frac{-1.7}{0.6261} \approx -2.7152$$

We can understand the -2.7152value to mean that assuming the truth of the null hypothesis the evidence that we collected from the sample mean is 2.7152standard deviations below the hypothesized population mean.

Since the alternative hypothesis contains the $<$ sign, this hypothesis test is a one-tailed test, and more extreme values would be values to the left, resulting in a left-tailed test. We have the following visualization for computing the p-value.

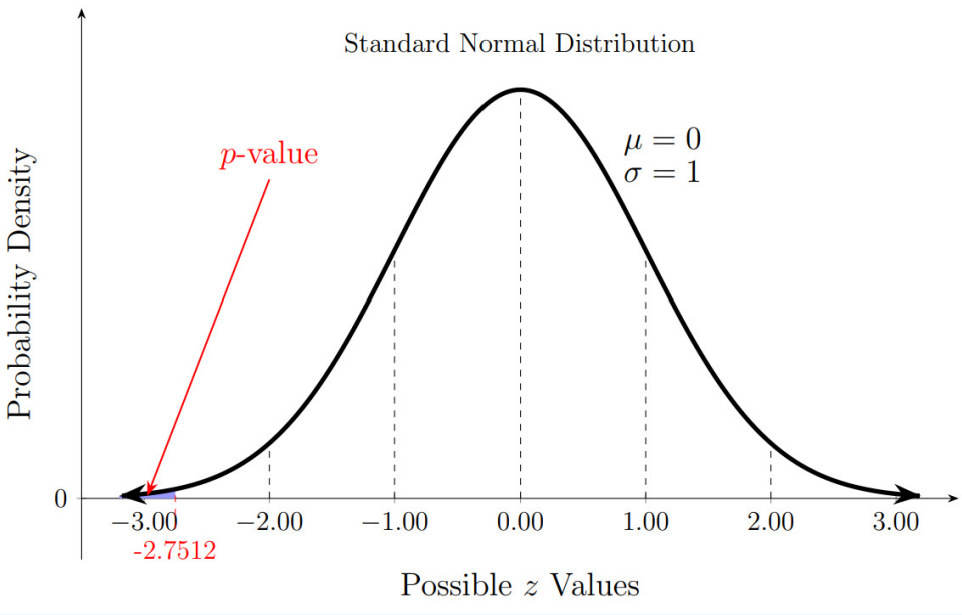


Figure 7.2.4 Standard normal distribution

$$p\text{-value} = \text{NORM.S.DIST}\left(\frac{-1.7}{\frac{4.2}{\sqrt{45}}}, 1\right) \approx \text{NORM.S.DIST}\left(-2.7152, 1\right) \approx 0.0033$$

We again produce the same p-value which yields that there is sufficient evidence to reject the null hypothesis in favor of concluding the alternative hypothesis: the genetically altered popular trees have less lignin naturally present.

Claims on Population Means (σ unknown)

We are now prepared to move to the more common situation: testing hypotheses about population means when the population standard deviation is unknown. The added complication comes from the fact that we do not know the standard deviation of the sampling distribution. We can estimate the population standard deviation using the sample standard deviation from our collected sample, but using this estimate has ramifications.

Recall constructing confidence intervals for population means when σ was unknown, we considered what happens under the t-transformation.

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}}$$

We concluded that the t variable followed a particular distribution, the Student's t-distribution with $n-1$ degrees of freedom. Notice how similar the formula for the variable t is to the formula for calculating the test statistic when σ is known. The only difference is that one formula has an s while the other has a σ. Just as the z-score transformation provided the formula to calculate the test statistic when σ is known, the t-transformation provides the formula we use to compute the test statistic when σ is unknown. We, therefore, know the distribution of test statistics when σ is unknown is the t-distribution with $n-1$ degrees of freedom. We use this fact to compute the p-value for testing claims on population means when σ is unknown. For a refresher on the t-distribution see Section 6.4. The two processes for testing hypotheses about population means are very similar. The main difference is the distribution in which we calculate the p-value. When σ is known, we use the standard normal distribution. When σ is unknown, we use the t-distribution with $n-1$ degrees of freedom.

Before we test hypotheses about population means with σ unknown, let us review the overall process of hypothesis testing to reinforce the procedure and highlight the distinctions between various situations.

1. Use natural observation, previous experimental results, or the claims of others to formulate a hypothesis that warrants testing. Within the context of means, each observation must admit some quantitative fact that can be measured and averaged. This

excludes considerations of whether or not observations have a particular quality; that will be studied in the section on claims on population proportions.

2. Identify a competing hypothesis and consider the ramifications of acting as if one of the hypotheses is true when, in fact, it is not. Name the hypothesis with the less drastic ramifications as the null hypothesis. The novel or claimed hypothesis is generally the alternative hypothesis. See note in the previous section regarding other ways to help distinguish between null and alternative hypotheses.
3. Determine the methodology of collecting evidence against the null hypothesis and determine what constitutes sufficient evidence by setting the level of significance. Make sure the design meets the requirements of the test intended to be conducted. For claims on population means, ensure that the sample is randomly selected and that either the underlying population is normally distributed or that the sample is large enough that the sampling distribution of sample means is approximately normal. In most cases, $n > 30$ will be sufficient.
4. Conduct the experiment and collect the evidence.
5. Compute the test statistic. Be sure to make the distinction between sample and population standard deviations. The most common situation is that we only have access to the sample standard deviation s and, therefore, must use the t-transformation to compute our test statistic. We often denote the test statistic based on which transformation is used. If the population standard deviation is known, the z-score transformation is used, and the test statistic is denoted with a z. If the population standard deviation is unknown, the t-transformation is used and the test statistic is denoted with a t.
6. Use the hypotheses to determine whether a test is a left-tailed, right-tailed, or two-tailed test. Note that the directions match the sign in the alternative hypothesis.
7. Determine the p-value by considering the test statistic, the appropriate distribution, and the type of test and then using technology to make an appropriate calculation.
8. Compare the p-value to the α value. If the p-value $< \alpha$ value, then we reject the null hypothesis in favor of the alternative hypothesis. If the p-value $\geq \alpha$ value, we fail to reject the null hypothesis.

? Text Exercise 7.2.4

A guest speaker at a local library presented on the change in human physical characteristics over the last two centuries in the United States. The presenter claimed that male height has consistently increased over that time and will continue to do so. The last evidence cited in this regard was in 1970, stating that the average height of adult males was 176 centimeters. Given the span of over 50 years, we decide to test the hypothesis that the average height of adult males in 2024 has increased since 1970.

Suppose we collect a sample of 16 adult males randomly selected and measure their heights in centimeters. The data is presented below. We decide to conduct the hypothesis test at an α value of 0.05.

$$169, 171, 171.5, 173, 173, 174.25, 174.5, 175, 176, 177, 177.75, 178, 178, 179, 179.5, 180$$

Answer

We must confirm that our circumstances enable a hypothesis test to be conducted; we need a random sample and reasonable confidence that the shape of the sampling distribution of sampling means is approximately normal. The first component, the random sample, is easily confirmed. The sample size chosen is 16, which does not meet the typical threshold of more than 30.. We must recall that male and female adult height are normally distributed. We, therefore, have that the sampling distribution of sample means is normally distributed. We can conduct the test. Note that no population standard deviations are given. We must utilize the recently discussed method that involves the t-distribution!

The guest speaker made the claim that the average height of adult males has increased over time. The population mean was 176 centimeters back in 1970; so, we identify one of the hypotheses as $\mu > 176$ centimeters. The opposite hypothesis would thus be $\mu \leq 176$ centimeters. This is a one-tailed test because both the average being the same or smaller than before are equally detrimental to the claims of the guest speaker. In both cases, the ramifications of acting as if one hypothesis is true when it is false seem to be mild. So, we pick the null hypothesis to be the one contrary to the claimed hypothesis. We settle on the hypotheses as follows.

$$H_0 : \mu \leq 176 \text{ centimeters}$$
$$H_1 : \mu > 176 \text{ centimeters}$$

Given the hypotheses, we have a right-tailed test. We compute our test statistic.

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

To do this, we must compute the sample mean and sample standard deviation using the values collected from our random sample. We produce the following results: $n = 16$, $\bar{x} \approx 175.4063$ centimeters, and $s \approx 3.2887$ centimeters.

At this stage, we notice that our sample mean is less than the hypothesized population mean. Since we have a right-tailed test, we are looking for evidence against the null hypothesis in the form of sample means larger than the hypothesized value. We can thus immediately conclude that there is not sufficient evidence to reject the null hypothesis. We will show the remainder of the computation to solidify the process and strengthen our conclusion.

$$t \approx \frac{175.4063 - 176}{\frac{3.2887}{\sqrt{16}}} \approx -0.7222$$

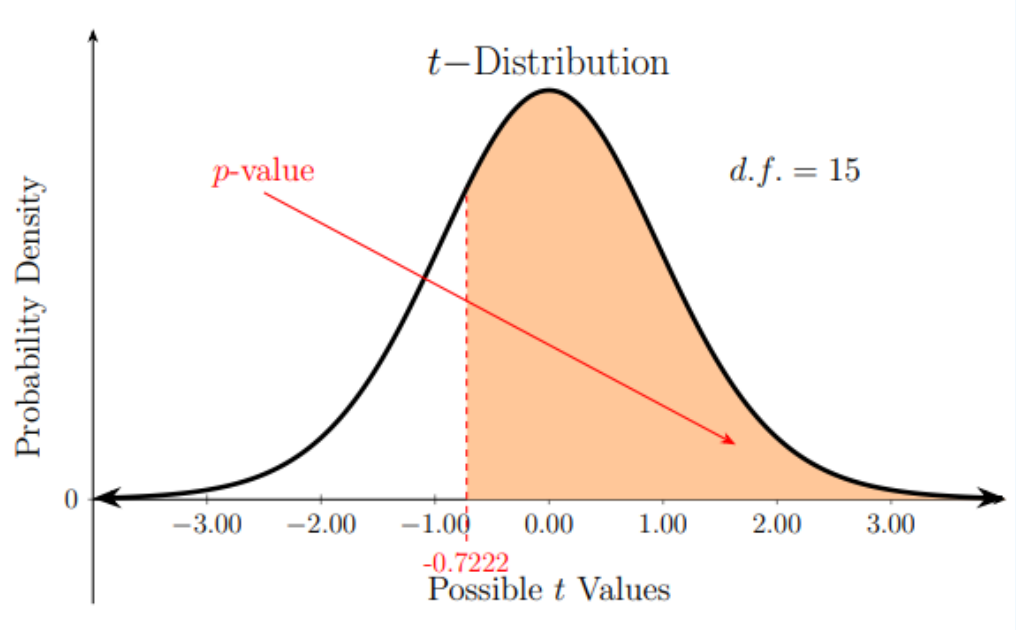


Figure 7.2.5 Right-tailed test with $t = -0.7222$ using t-distribution with 15 degrees of freedom

As we can see, the shaded area is over half of the area due to the symmetry of the t-distribution. We now compute the p-value using technology.

$$p\text{-value} \approx 1 - \text{T.DIST}(-0.7222, 15, 1) \approx 1 - 0.2406 \approx 0.7594$$

The p-value is larger than the α value of 0.05 We conclude the test by failing to reject the null hypothesis. There is not sufficient evidence to support the guest speaker's claim.

It is still possible that the average height of adult males has increased, and something rare occurred in the act of sampling. It is also possible that the average height is the same as it was. Based on the evidence, we may also be open to the idea that the average height may be smaller. Reviewing the guest speaker's reasoning behind why the heights have been increasing may be prudent. Is there a faulty assumption? Is there an explanation as to why it may have been increasing and now possibly decreasing? Perhaps we can conduct another random experiment to test whether the mean is now less than it was before. Hypothesis tests that fail to reject the null hypothesis can still inform further research and inquiry.

? Text Exercise 7.2.5

A study published in 2021, concluded that the average weekly recreational screen time of $18 - 29$ year olds (emerging adults) increased from 2018 to 2020 during the pandemic estimating the average weekly recreational screen time with the confidence interval 28.5 ± 11.6 hours. Recreational screen time does not include screen time associated with work or school.

The pandemic is now behind us, but the effects of the pandemic are still playing out. A researcher is still interested in the weekly recreational screen time of emerging adults and conducts a study on 56 randomly selected emerging adults with a sample mean of 30.4 hours and a sample standard deviation of 6.3 hours. The researcher adopts the following hypotheses to be tested at the 0.05 level of significance. Conduct the test.

$$H_0 : \mu = 28.5 \text{ hours}$$
$$H_1 : \mu \neq 28.5 \text{ hours}$$

Note that this sample data is completely fabricated.

Answer

Having been given the hypotheses, we note that we are conducting a two-tailed test. The researcher adopted the central value of confidence interval from the study to compare the current data. We do not know the population standard deviation and thus operate within the realm of the t-transformation and t-distribution.

$$t = \frac{30.4 - 28.5}{\frac{6.3}{\sqrt{56}}} \approx 2.2569$$

Since we have a two-tailed test, we determine the test-statistic that is equally as extreme as the computed test statistic in the opposite direction. Again due to symmetry, this is the value of equal magnitude but opposite sign. We have the following visualization for computing the p-value.

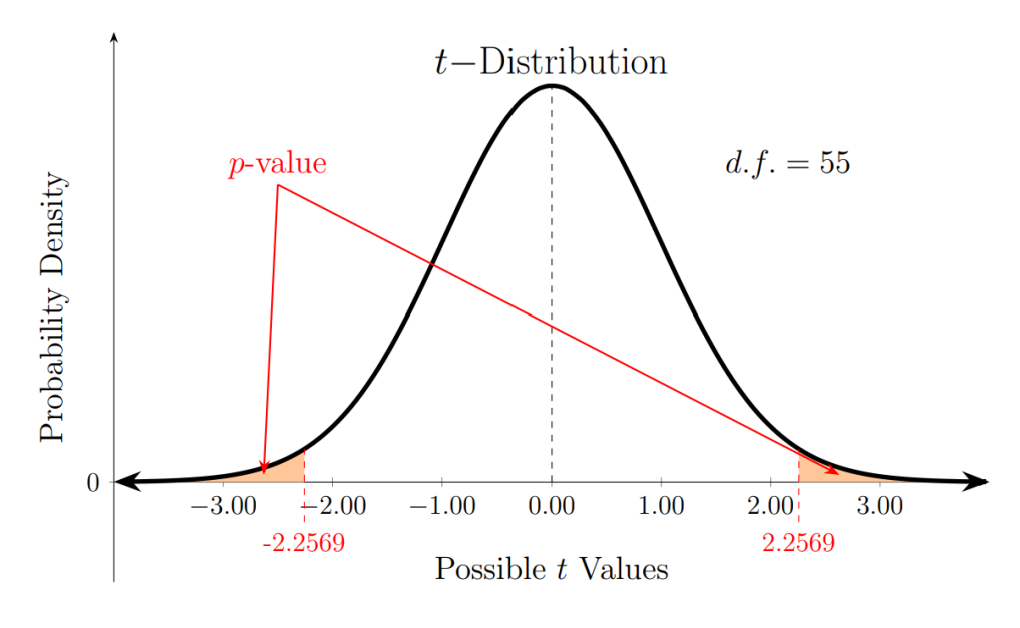


Figure $7.2.6$ Two-tailed test with $t = \pm 2.2569$ using t-distribution with 55 degrees of freedom

$$p\text{-value} = 2 \cdot \text{T.DIST}(-2.2569, 55, 1) \approx 2 \cdot 0.0140 \approx 0.0280$$

We again used symmetry, noting that the boundaries of the tails are equidistant from 0, so that we can double the area found in one of the tails. The left tail can be computed directly using the negative value of the two test statistics.

We compare 0.0280 with 0.05 and find that $0.0280 < 0.05$. We have sufficient evidence to reject the null hypothesis that the average weekly recreational screen time for emerging adults in 2024 is 28.5 hours and conclude that the average weekly recreational screen time for emerging adults in 2024 is not 28.5 hours. The evidence indicates that the average may actually be higher, but to reach such a conclusion, another study must be conducted.

7.3: Claims on Dependent Paired Variables

Learning Objectives

- Distinguish between dependent and independent samples
- Develop and apply hypothesis testing for dependent paired variables

Section 7.3 Excel File: (contains all of the data sets for this section)

Review and Preview

Recall the two studies about weekly recreational screen time (one real and one fabricated) from Text Exercise 7.2.5. It might be tempting to conclude that the average weekly recreational screen time for emerging adults in 2024 is different from the average weekly recreational screen time for emerging adults in 2020 since we have evidence to say that the average in 2024 is different from 28.5 hours, but we must exercise a little caution. The original study (the real one) reported an estimate for the population mean using a confidence interval with a margin of error of 11.6 hours. This is a large margin of error. The average weekly recreational screen time for emerging adults in 2020 could be anywhere from 16.9 hours to 40.1 hours. We do not know precisely where it falls. So, using 28.5 hours as the conclusive average for 2020 is questionable. We will always be using estimates of parameters unless we conduct a census on the population. One might ask how can we ever proceed with these sorts of comparisons? Did not the standard population means from other problems come from interval estimates as well? The short answer is yes, they did, but there is more at play.

We first note that we can control the size of the margin of error by balancing confidence level and sample size. A more precise estimate can be obtained using a larger sample. If the margin of error were only 0.1 hours, we might feel more confident in treating the population mean as 28.5. We can also approach the problem using a different frame of reference. The general idea is that we are comparing two populations so we should make comparisons using the data from both populations. We compare them by collecting a random sample from each population and then analyzing the differences in the samples.

One methodology compared recreational screen time in 2018 to 2020. The original study used data from 2018 and 2020 from the same set of people. The researchers could study the difference in recreational screen time by each member of the sample. They had one sample from 2018 and another sample from 2020, but they were dependent upon each other because they consisted of the same set of people. We describe such a situation as one with dependent samples. Tests using dependent samples, often referred to as tests on dependent paired variables, provide strong results because they reduce the influence of confounding variables; there is less variation across one subject as a single treatment is applied than the variation present across the members of the population, but this is not the only way two populations can be compared.

Imagine the difficulty of keeping track of hundreds of participants over the course of months or years. Is it possible to make comparisons between 2018 and 2020 without conducting such a longitudinal study? The answer is yes. Two samples can be taken independently of each other. A random sample may be taken from one population and then another random sample may be taken from the other population. In the context of recreational screen time, a random sample may be taken in 2018 and then another random sample may be taken in 2020. Here we are not guaranteed that the same people will be in the two samples. It is possible that there is overlap, but the fact that a person was in the first sample does not affect the probability that they are in the second sample. We describe such a situation as one with independent samples. We will not address the methodologies involved in such claims in this text, but the interested reader is encouraged to study it independently. We now begin our development of testing claims on dependent paired variables.

Claims on Dependent Paired Variables

Researchers in medicine, education, and business are often interested in studying the effect of some treatment, educational practice, or product. It is quite natural to assess the patient, student, or consumer prior to some treatment and then assess them once the treatment has been in effect. Consider medical research: doctors can conduct pre-assessments and post-assessments to gauge the impact of a particular medical intervention on a random sample of patients. The doctors could simply compare the pre-assessment and post-assessment averages as if the samples were independently gathered, but there is a connection between the samples that is not being acknowledged, namely, that the same patients have two assessment values. We can measure effect of the medical intervention on each patient by considering the difference in the pre-assessment and post-assessment. In studying these paired differences, it is like we are studying a single sample and can utilize techniques already developed in this chapter to test claims.

A common concern of many people, especially in the medical community, is the consumption of chicken eggs. Previous research seems to indicate the possibility of a tie to heart disease and diabetes, but studies require independent attempts at reproducing the same results to verify that they weren't produced by chance. Suppose a medical researcher designs and conducts the following study to test the impact eating 2 chicken eggs a day has on LDL cholesterol (low-density lipoprotein cholesterol (the bad cholesterol)) levels in the body.

Given the varying conclusions of the previous medical research and the number of confounding variables that cloud their results, this particular researcher decides to test whether or not there is any effect in adopting the consumption of 2 chicken eggs a day and will test at a significance level of 0.05.

Participants are randomly sampled from the population at large. Each participant is asked to abstain from eating chicken eggs for the span of 3 months to normalize the sample to a diet without chicken eggs. Each participant's LDL cholesterol is measured the morning after completing the 3 month normalization period and is expected to have been fasting from midnight the night before. The participants then eat 2 chicken eggs scrambled using a teaspoon of olive oil each day for breakfast for an entire month. Participants are expected to maintain their regular diet otherwise. At the end of a month, participants again have their LDL cholesterol measured in the same fashion as before.

Table 7.3.1: Initial and Final LDL Cholesterol Readings

Participant #	Initial LDL (mg/dL)	Final LDL (mg/dL)
1	189	189
2	110	101
3	155	158
4	97	94
5	83	73
6	75	73
7	182	189
8	177	180
9	160	151
10	185	184
11	72	72
12	169	171
13	87	86
14	112	118
15	112	118
16	107	104
17	168	174
18	190	194
19	120	126
20	122	125
21	175	167
22	168	178
23	106	104

Participant #	Initial LDL (mg/dL)	Final LDL (mg/dL)
24	108	110
25	93	99
26	129	139
27	95	94
28	63	68
29	176	170
30	186	191
31	171	175
32	154	154
33	78	76
34	156	164
35	170	160

To analyze the results of this hypothetical medical study (the results were fabricated for the purposes of the book), we treat the two samples as dependent samples given that the variables of interest (LDL cholesterol levels before and after) can be matched by participant. We are interested in the change in cholesterol level after having the medical intervention of eating 2 scrambled chicken eggs a day for a month. To compute the change, we will need to compute the difference between the final measurement and the initial measurement, Final LDL $-$ Initial LDL. A positive difference indicates that the LDL level increased; while, a negative difference indicates that the LDL level decreased. We will conduct our analyses on the values of these differences. To emphasize the fact that we are studying the differences of dependent paired variables, we will utilize the following notation for means and standard deviations: μ_d, $\bar{x}_d$, σ_d, and s_d.

With this notation in hand, let us formulate our hypotheses regarding the average value of these differences. The researcher wants to determine whether eating 2 chicken eggs a day has any effect on LDL levels. This would be an increase or decrease. If there is no effect, the average of the differences will be 0. If there is an effect, the average of the differences will not be 0. We adopt the former as our null hypothesis because chicken eggs are a relatively cheap source of protein and other nutrients that have been consumed consistently in larger quantities for a long time.

$$H_0 : \mu_d = 0 \text{ mg/dL}$$
$$H_1 : \mu_d \neq 0 \text{ mg/dL}$$

Having our hypotheses in hand, we compute the differences to analyze and ensure that we met the requirements necessary to conduct the hypothesis test. We have a random sample with a sample of 35 participants. Just like in our previous tests, we need either that the underlying distribution, the distribution of all these differences, is normal or that the sample is large enough for the Central Limit Theorem to assert that the sampling distribution of sample means is approximately normal. Since $n = 35$, we will proceed using the latter as our justification.

Table 7.3.2: Initial and Final LDL Cholesterol Readings with Differences

Participant #	Initial LDL (mg/dL)	Final LDL (mg/dL)	Difference (Final - Initial) (mg\dL)
1	189	189	0
2	110	101	-9
3	155	158	3
4	97	94	-3

Participant #	Initial LDL (mg/dL)	Final LDL (mg/dL)	Difference (Final - Initial) (mg\dL)
5	83	73	-10
6	75	73	-2
7	182	189	7
8	177	180	3
9	160	151	-9
10	185	184	-1
11	72	72	0
12	169	171	2
13	87	86	-1
14	112	118	6
15	112	118	6
16	107	104	-3
17	168	174	6
18	190	194	4
19	120	126	6
20	122	125	3
21	175	167	-8
22	168	178	10
23	106	104	-2
24	108	110	2
25	93	99	6
26	129	139	10
27	95	94	-1
28	63	68	5
29	176	170	-6
30	186	191	5
31	171	175	4
32	154	154	0
33	78	76	-2
34	156	164	8
35	170	160	-10

We do not know anything about the population parameters, so we will have to conduct our test using the t-transformation test statistic; hypotheses tests on dependent paired variables in this context are often referred to as a paired t-tests. We compute the sample mean and standard deviation using the difference values in the fourth column, compute the test statistic under the

assumption that the null hypothesis is true, and produce a visualization for computing the p-value. Notice that since the fourth column has 35 data points, we will use $n = 35$ in our computations, despite the fact that we recorded 70 values in total. $\bar{x}_d \approx 0.8286\,\text{mg\dL}$. $s_d \approx 5.6386\,\text{mg\dL}$.

$$t \approx \frac{0.8286 - 0}{\frac{5.6386}{\sqrt{35}}} \approx 0.8694.$$

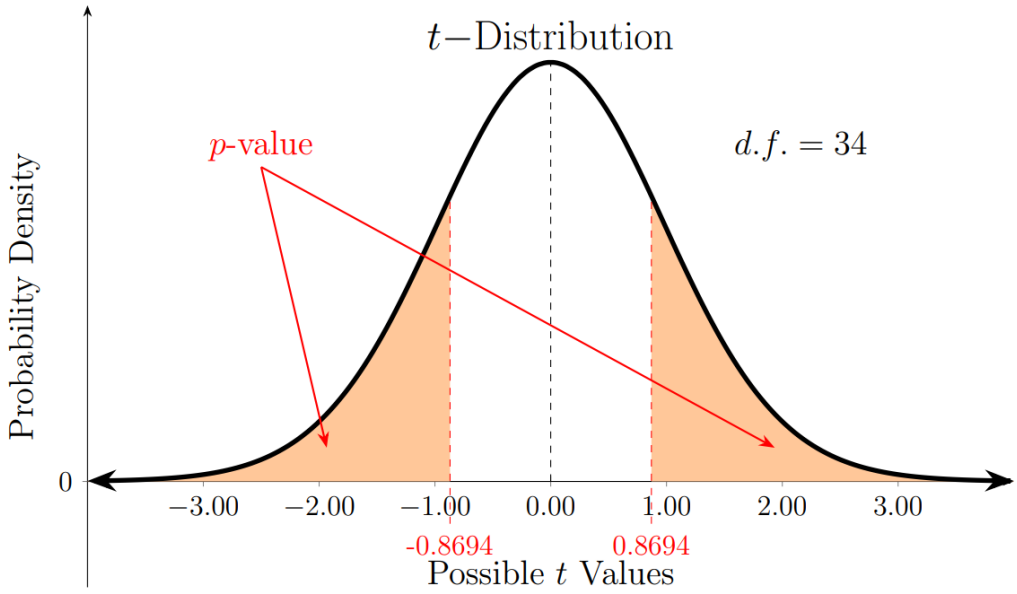


Figure 7.3.1: t-distribution for LDL Cholesterol Readings

$$p\text{-value} \approx 2 \cdot \text{T.DIST}(-0.8694, 34, 1) \approx 2 \cdot 0.1954 \approx 0.3908$$

Given that the p-value is greater than the α value, we fail to reject the null hypothesis. There is not sufficient evidence to say that eating 2 chicken eggs per day in the manner specified in the study alters the amount of LDL cholesterol in one's system over the course of a month.

? Text Exercise 7.3.1

An athletic training company executive officer recently discovered the knees-over-toes guy, a trainer with a seemingly effective approach to living well through exercise focused on whole body movement, flexibility, and overall strength. The trainer claims that his approach helps people dunk basketballs. As this is an area of strategic growth for his company, the executive officer was enticed and decided to test the strategy on his basketball clients for a year to assess the growth in the height of the clients vertical jump. A random sample of 31 male clients was selected to participate in the study. Initial and final vertical jumps were measured in inches (see table below). Conduct the hypothesis test at a 0.02 significance level. Note that program is real, but this study is fabricated for the purposes of the book.

Table 7.3.3: Initial and Final Jump Height in Inches

Client #	Initial Jump Height (in)	Final Jump Height (in)
1	18	25
2	22	24
3	24	25
4	16	17
5	18	24

Client #	Initial Jump Height (in)	Final Jump Height (in)
6	22	24
7	24	28
8	26	29
9	15	20
10	18	19
11	14	20
12	16	16
13	24	31
14	16	21
15	24	32
16	24	26
17	24	25
18	14	22
19	18	20
20	23	29
21	15	22
22	18	19
23	25	32
24	18	23
25	17	20
26	16	19
27	22	23
28	20	22
29	25	29
30	24	24
31	14	14

Answer

We treat the two samples as dependent samples given that the variables of interest (vertical jump height) came from the same participant pool and we can match the values by participant. We are again interested in the change in the variable of interest after having some intervention; in this case, the intervention is a particular form of athletic training. To compute the change, we will need to compute the difference between the final measurement and the initial measurement. Again, a positive difference indicates that the intervention increased the jump height; while, a negative difference indicates that the jump height decreased. We will again conduct our analyses on the values of these differences.

The company officer will only be interested in the new program if it increases clients' jump heights. Increasing the jump height would result in a positive difference on average. The company officer does not want to assume that the program is effective without evidence; we, therefore, have the following hypotheses for our test.

$$H_0 : \mu_d \leq 0 \text{ in}$$
$$H_1 : \mu_d > 0 \text{ in}$$

Since the study used a random sample of 31 clients, the hypothesis test can be conducted. We compute the differences in the following table.

Table 7.3.4 Initial and Final Jump Height with Differences in Inches

Client #	Initial Jump Height (in)	Final Jump Height (in)	Difference (Final - Initial) (in)
1	18	25	7
2	22	24	2
3	24	25	1
4	16	17	1
5	18	24	6
6	22	24	2
7	24	28	4
8	26	29	3
9	15	20	5
10	18	19	1
11	14	20	6
12	16	16	0
13	24	31	7
14	16	21	5
15	24	32	8
16	24	26	2
17	24	25	1
18	14	22	8
19	18	20	2
20	23	29	6
21	15	22	7
22	18	19	1
23	25	32	7
24	18	23	5
25	17	20	3
26	16	19	3
27	22	23	1
28	20	22	2
29	25	29	4

Client #	Initial Jump Height (in)	Final Jump Height (in)	Difference (Final - Initial) (in)
30	24	24	0
31	14	14	0

We will again conduct a paired t-test. $\bar{x}_d \approx 3.5484$inches. $s_d \approx 2.5928$inches.

$$t \approx \frac{3.5484 - 0}{\frac{2.5928}{\sqrt{31}}} \approx 7.6198$$

.

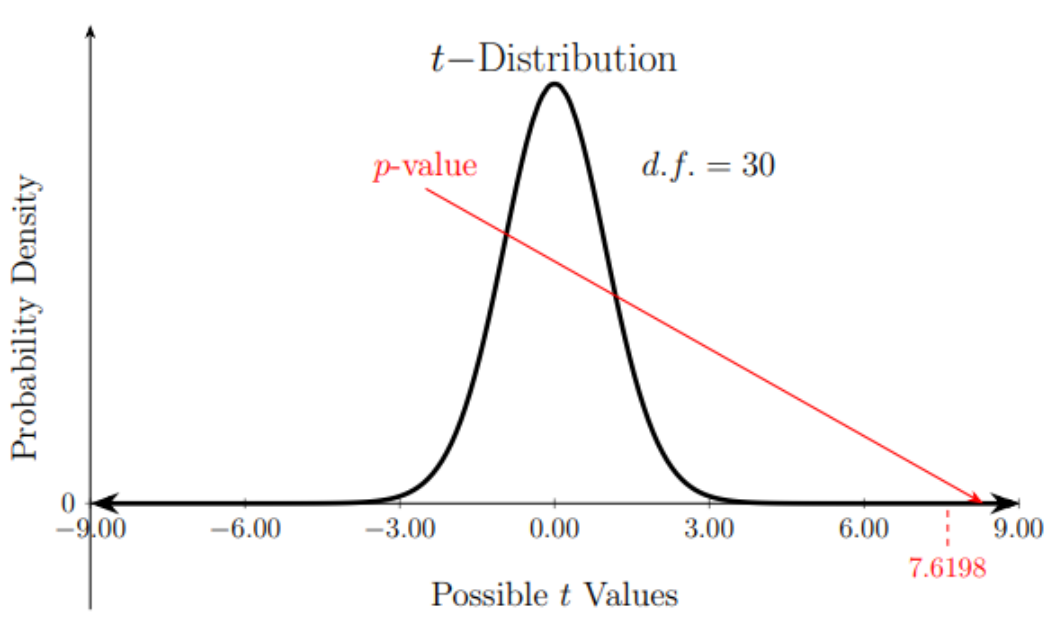


Figure 7.3.2 Right-tailed test with $t = 7.6198$

$$p\text{-value} \approx 1 - \text{T.DIST}(7.6198, 30, 1) \approx 8.4615 \cdot 10^{-9}$$

Given that the p-value is less than the α value, we reject the null hypothesis. There is sufficient evidence to say that over the course of a year using the knees-over-toes guy's training regiment the average height of clients' vertical jumps increased.

? Text Exercise 7.3.2

Many people have been concerned with carbon emissions from automobiles. Various governments have enacted policies that set emission standards and goals for new cars. A government is giving automobile manufacturers 10 years to reach the emission standards, but each year the manufacturers have to show that progress has been made by reducing carbon dioxide emissions across updated models within each class of vehicles in the amount of at least 10 grams of carbon dioxide per mile driven.

An automobile manufacturer's analyses indicate that they will not meet the emission progression threshold for their four-door sedans. They are aware of certain studies that state that the fuels with higher ethanol concentrations produce less emissions. It happens that the motors in this class of cars work well with pure gasoline and gasoline blended with ethanol. Without the time to redesign enough models to meet the progression requirements, the company considers selling their four-door sedans as requiring gasoline blended with a high ethanol concentration. They are hoping the difference from the fuel will be enough to satisfy the requirements. With all the varieties in models, the company makes over 600 different four-door sedans. They

randomly select 31 models to test the carbon dioxide emissions and then compare the results to the results of the previous year. The results are presented in the table below. Test the hypothesis at the 0.05 significance level.

Table 7.3.5: Four-Door Sedan Emissions

Model of Four-Door Sedan #	Emissions from Last Year (g/mi)	Emissions from This Year with Blend (g/mi)
1	483	471
2	468	456
3	409	401
4	457	452
5	461	447
6	403	396
7	408	396
8	414	398
9	429	422
10	443	428
11	467	460
12	386	369
13	350	343
14	396	381
15	476	461
16	363	347
17	465	453
18	398	392
19	426	417
20	489	472
21	454	444
22	449	442
23	400	387
24	380	365
25	383	378
26	371	357
27	423	406
28	437	423
29	379	374
30	351	338
31	397	385

Answer

We treat the two samples as dependent samples given that the variables of interest (carbon dioxide emissions per mile driven) are paired by particular models of four-door sedans. We are again interested in the change in the variable of interest after having some intervention; in this case, blended fuel. To compute the change, we will need to compute the difference between the final measurement and the initial measurement. Again, a positive difference indicates that the intervention increased emission rates; while, a negative difference indicates a decrease in emission rates. We will again conduct our analyses on the values of these differences.

The company will only be interested if switching fuel specifications decreases carbon dioxide emission by at least 10 grams per mile driven on average. The company does not want to assume that this is the case without evidence. We form the following hypotheses.

$$H_0 : \mu_d \geq -10 \text{ g/mi}$$
$$H_1 : \mu_d < -10 \text{ g/mi}$$

Since the study used a random sample of 31 models of four-door sedans, the hypothesis test can be conducted. We compute the differences in the following table.

Table 7.3.6 Four-Door Sedan Emissions with Differences

Model of Four-Door Sedan #	Emissions from Last Year (g/mi)	Emissions from This year with Blend (g/mi)	Difference (g/mi)
1	483	471	−12
2	468	456	−12
3	409	401	−8
4	457	452	−5
5	461	447	−14
6	403	396	−7
7	408	396	−12
8	414	398	−16
9	429	422	−7
10	443	428	−15
11	467	460	−7
12	386	369	−17
13	350	343	−7
14	396	381	−15
15	476	461	−15
16	363	347	−16
17	465	453	−12
18	398	392	−6
19	426	417	−9
20	489	472	−17
21	454	444	−10
22	449	442	−7

Model of Four-Door Sedan #	Emissions from Last Year (g/mi)	Emissions from This year with Blend (g/mi)	Difference (g/mi)
23	400	387	-13
24	380	365	-15
25	383	378	-5
26	371	357	-14
27	423	406	-17
28	437	423	-14
29	379	374	-5
30	351	338	-13
31	397	385	-12

We will again conduct a paired t-test. $\bar{x}_d \approx -11.4194$g/mi. $s_d \approx 4.0148$g/mi.

$$t \approx \frac{-11.4194-(-10)}{\frac{4.0148}{\sqrt{31}}} \approx -1.9684$$

.

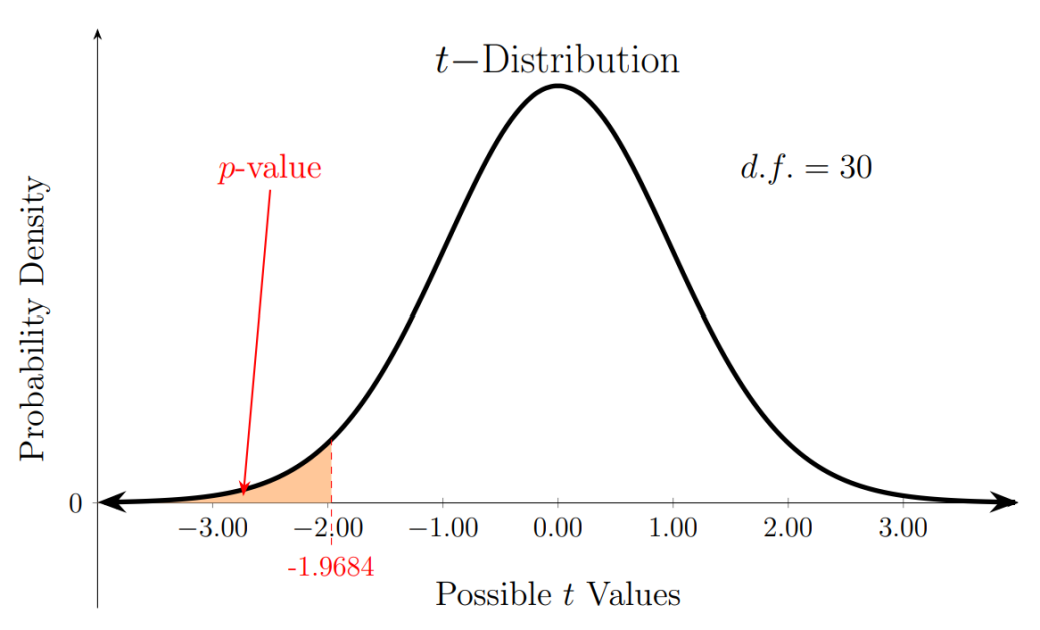


Figure 7.3.3 Left-tailed test with $t = -1.9684$

p-value $\approx \text{T.DIST}(-1.9684, 30, 1) \approx 0.0292$

Given that the p-value is less than the α value, we reject the null hypothesis. There is sufficient evidence to say that switching the fuel classification of the company's four-door sedans to requiring a ethanol-gasoline blended fuel with high concentrations of ethanol will allow the company to meet the emission progress standards set by the government. The progress may not reflect the intent of the law but seems to pass the letter of the law.

7.4: Claims on Population Proportions

Learning Objectives

- Conduct hypothesis testing on claims regarding population proportions using the sampling distribution of sample proportions
- Conduct hypothesis testing on claims regarding population proportions using test statistics

Review and Preview

Recall that proportions measure the percentage of observations that admit a certain quality. We might be interested in the percentage of the population (who are registered to vote) that will actually vote in an upcoming election. Each registered voter either will vote or will not vote; the registered voter either has the quality or does not have the quality. Remember that we denote the proportion of the population with a quality of interest using p. Since there are only two states regarding the quality, everybody else does not have the quality; we denote this proportion with q. The entire population is covered between the observations with the quality and those without the quality; we thus know that $p+q=1$.

We may be even more interested in the proportion of registered voters who will vote for a particular candidate or a particular item on the ballot. In certain cases, particular proportions of affirmative votes are required for an item to pass. Can we test that there is enough support for a particular candidate or item to pass before the election occurs or before all of the ballots are counted? These questions center on claims about population proportions and is the topic of this section.

Testing claims on population proportions intimately involves the sampling distribution of sample proportions. In order to compute p-values, we need to know the approximate shape of the sampling distribution. We have seen and utilized the fact that the sampling distribution of sample proportions is approximately normal with $\mu_{\hat{p}}=p$ and $\sigma_{\hat{p}}=\sqrt{\frac{pq}{n}}$ when our sample size n is large enough that we expect more than 5 observations with the quality and more than 5 observations without the quality to be in our sample. To check this condition, we checked that the following two inequalities were satisfied: $np>5$ and $nq>5$. With such a preview and having several sections of hypothesis testing under our belts, let us begin testing claims on population proportions.

Testing Claims on Population Proportions

When conducting hypothesis testing, we do not know the value of the population proportion p, but this is okay because we compute the p-value under the assumption that the null hypothesis is true. We will thus be operating under the assumption that the population proportion is equal to some particular value which we denote as p_0. This notation leads to q_0 which is the hypothesized proportion of the population without the quality. So, in order to conduct hypothesis testing on claims about population proportions, we need our samples to be randomly chosen and of such a size that $np_0>5$ and $nq_0>5$. When these conditions are met, we can conduct the probability assessment using a normal distribution with $\mu_{\hat{p}}=p_0$ and $\sigma_{\hat{p}}=\sqrt{\frac{p_0q_0}{n}}$.

? Text Exercise 7.4.1

The success of a manufacturing plant that produces tens of thousands of motion-detecting sensors each week requires a high degree of quality assurance and quality control. As such, the plant sets the standard that at most 2.5% of the sensors produced at the plant will be defective. To test that the plant is meeting its production standards, random samples of 500 sensors are taken each week and tested. The company tests at the 0.03 level of significance. Last week, the sample contained 20 defective sensors.

1. Conduct the hypothesis test using the sampling distribution of sampling proportions and interpret the conclusions within the context of the problem.

Answer

We are considering a claim on population proportions because we are considering the percentage of sensors that have the quality that they are defective. The company set the standards that $p<0.025$. This forms one of our hypotheses. The competing hypothesis would thus be that $p\geq 0.025$. The company does not want to have the default position that the machinery is not working; otherwise, they will frequently be conducting unnecessary maintenance.

$$H_0 : p \leq 0.025$$
$$H_1 : p > 0.025$$

We note that under the assumption that the null hypothesis is true, the largest p-value will be computed when the value is assumed to be 0.025. We thus set $p_0 = 0.05$ to ensure the conditions for the test are met. Noting that $q_0 = 1 - p_0 = 0.95$ and $n = 500$, we have $np_0 = 12.5$ and $nq_0 = 487.5$. With the two inequalities met and the sample being randomly chosen, we can conduct the hypothesis test.

Under the assumption that the null hypothesis is true and in the situation that produces the largest p-value, we have $\mu_{\hat{p}} = 0.025$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.025 \cdot 0.975}{500}} \approx 0.007$. The random sample of sensors from last week had 20 defective sensors. This is not the sample proportion. Proportions fall inclusively between 0 and 1. The proportion of defective sensors in the sample is the percent of defective sensors in the sample; thus, $\hat{p} = \frac{20}{500} = 0.04$. Given the hypotheses, we have a right-tailed test and thus visualize the test in the following figure.

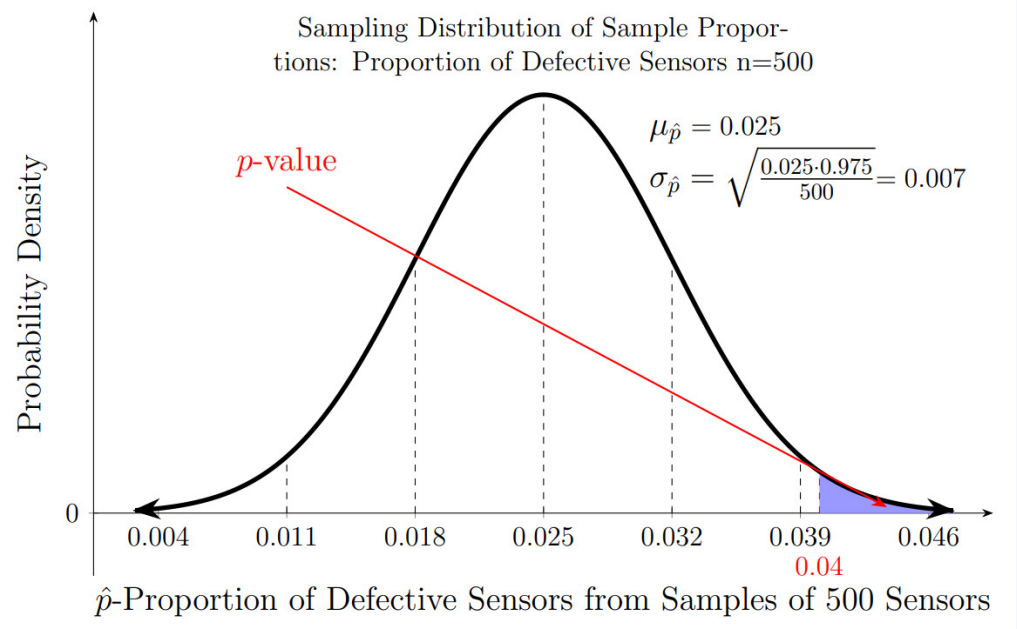


Figure 7.4.1: Sampling distribution of sample proportions

$$p\text{-value} \approx 1 - \text{NORM.DIST}(0.04, 0.025, 0.007, 1) \approx 0.0158$$

Since the p-value is smaller than the α value, we have sufficient evidence to reject the null hypothesis that proportion of defective sensors is within the limit given the quality control standards set by the company. As such further investigation should happen regarding the sensors produced last week and the machinery should be checked before continuing production.

2. Determine a transformation that takes the sampling distribution of sample proportions to a common distribution and thus determine the formula for the test statistic within the context of hypothesis testing with claims on population proportions. Verify that your solution is correct by applying it in the context of this text exercise and obtaining the same p-value.

Answer

Since the sampling distribution of sample means is approximately normal when we are able to conduct hypothesis testing on claims about population parameters and we know the mean and the standard deviation, we can use the z-score transformation to map the sampling distribution to the standard normal distribution. We can thus define our test statistic as follows

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

We now apply this formula to the context of the text exercise to obtain a little validation. Using the values computed in the previous part, we have $n = 500, \mu_{\hat{p}} = 0.025, \sigma_{\hat{p}} = \sqrt{\frac{0.025 \cdot 0.975}{500}} \approx 0.007$, and $\hat{p} = \frac{20}{500} = 0.04$.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \approx \frac{0.04 - 0.025}{\sqrt{\frac{0.025 \cdot 0.975}{500}}} \approx 2.1483$$

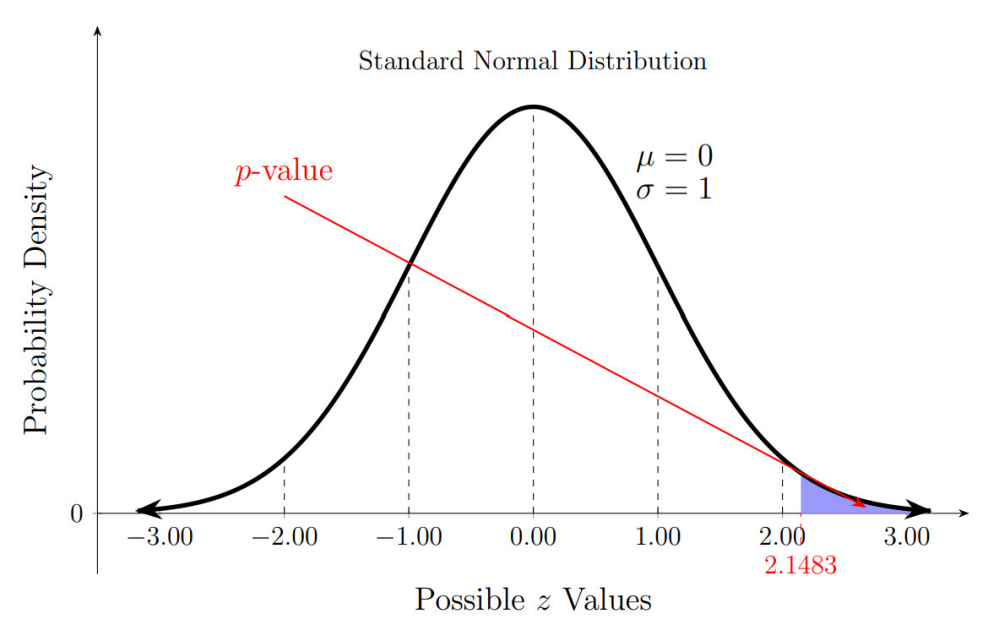


Figure 7.4.2 Standard normal distribution

$$p\text{-value} \approx 1 - \text{NORM.S.DIST}(2.1483, 1) \approx 0.0158$$

This produces the same p-value as computed from the sampling distribution of sample proportions in the previous part of the question. Indeed, we have settled on the proper formulation of test statistics in the realm of testing hypotheses on population proportions.

? Text Exercise 7.4.2

A large, public corporation with thousands of shareholders is considering purchasing another large corporation, but according to the bylaws by which the corporation was founded, to do so requires a two-thirds majority of shareholders to be in support of such a purchase. The chief operating officer is vehemently opposed to the acquisition and has been rallying the shareholders to vote against the purchase. The chief operating officer gets to set the agenda for the upcoming shareholder meeting and is trying to decide if the vote regarding the purchase should be held or postponed.

To facilitate this decision, the chief operating officer randomly selects 60 shareholders and has the human resources department contact them to assess their positions regarding the possible acquisition. After conducting these 60 conversations, the human resources department returns that 35 of the shareholders are planning to vote in favor of the acquisition. Conduct a hypothesis test from the perspective of the chief operating officer at the 0.05 significance level and decide, from his perspective, whether or not to schedule the vote during the upcoming meeting.

Answer

For the vote to pass, a two-thirds majority of shareholders need to vote in favor of the acquisition in order for it to pass. This fact determines our two hypotheses: $p \geq \frac{2}{3}$ and $p < \frac{2}{3}$.

The chief operating officer does not want the acquisition to pass and, therefore, wants the second hypothesis to be true. He has the control over when the vote occurs. He does not want to assume his position is going to win out. We thus set the hypotheses as follows.

$$H_0 : p \geq \frac{2}{3}$$
$$H_1 : p < \frac{2}{3}$$

We note that under the assumption that the null hypothesis is true, the largest p-value will be computed when the value is assumed to be $\frac{2}{3}$. We thus set $p_0 = \frac{2}{3}$ and note that $q_0 = \frac{1}{3}$ and $n = 500$. Thus we have $np_0 = 40$ and $nq_0 = 20$. With the two inequalities met and the sample being randomly chosen, we can conduct the hypothesis test.

We must determine the mean and standard deviation of the sampling distribution and the sample proportion in order to compute the test statistic and then compute the probability for this left-tailed test. $\mu_{\hat{p}} = \frac{2}{3}$, $\sigma_{\hat{p}} = \sqrt{\frac{\frac{2}{3} \cdot \frac{1}{3}}{60}} \approx 0.0609$, and $\hat{p} = \frac{35}{60} = \frac{7}{12} = 0.58\bar{3}$.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{\frac{7}{12} - \frac{2}{3}}{\sqrt{\frac{\frac{2}{3} \cdot \frac{1}{3}}{60}}} \approx -1.3693$$

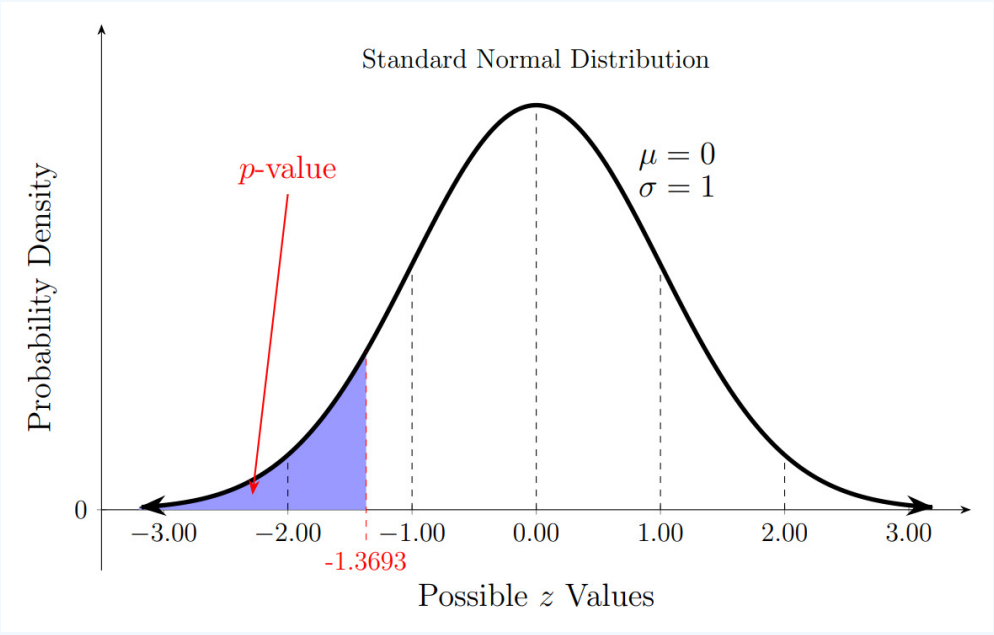


Figure 7.4.3 Standard normal distribution

$$p\text{-value} \approx \text{NORM.S.DIST}(-1.3693, 1) \approx 0.0855$$

The p-value is larger than the α value. There is not sufficient evidence to reject the null hypothesis that there is enough support for the acquisition to pass. Since the chief operating officer cannot feel confident that the acquisition is going to fail, he should postpone the vote in order to have more time to convince shareholders of his point of view.

? Text Exercise 7.4.3

We often assume that the probability of having a female baby is 50%, but there is mounting evidence that indicates this assumption does not align with reality. The Centers for Disease Control (CDC) of the United States keeps track of birth records and makes the data accessible to the public. In 2019, there were $3,747,540$ births in the United States with $1,830,094$ of those being female. This indicates that only 48.8345% of babies born in the United States in 2019 were female. In 2023, there were $3,519017$ births with $1,756,223$ being females which again produces a proportion of 48.8380% of babies being female. What about on a global scale?

The United Nations maintains records and an organization called Our World in Data maintains an article addressing the gender ratio. It is the general trend throughout history, at least for the last century, that more males are born than females globally. A study cited by Our World in Data indicates that the proportion of females at the time of conception is indeed 50%, which implies that the difference is caused by events occurring during pregnancies. An interested reader is encouraged to examine the article linked above.

Suppose we took a random sample of newly born infants from across the world and 5460 of them were female while 5733 of them were male. Would this constitute significant evidence against the common assumption that 50% of babies born are female? Test the claim at a significance level of 0.01.

Answer

The wording of the problem "evidence against the common assumption that 50% of babies born are female" indicates the null hypothesis. We thus have a two-tailed test with the following hypotheses.

$$H_0 : p = 0.50$$
$$H_1 : p \neq 0.50$$

In order to confirm the requirements for the test, we need to compute the sample size $n = 5460 + 5733 = 11193$. Since $p_0 = q_0$, we have only one inequality to check. Half of $11193$is much more than 5; so, we have the requirements met. Our sample is large enough and was randomly selected.

$$\mu_{\hat{p}} = 0.50, \sigma_{\hat{p}} = \sqrt{\frac{0.5 \cdot 0.5}{11193}} \approx 0.0047, \text{and } \hat{p} = \frac{5460}{11193} \approx 0.4878.$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \approx \frac{0.4878 - 0.50}{0.0047} \approx -2.5804$$

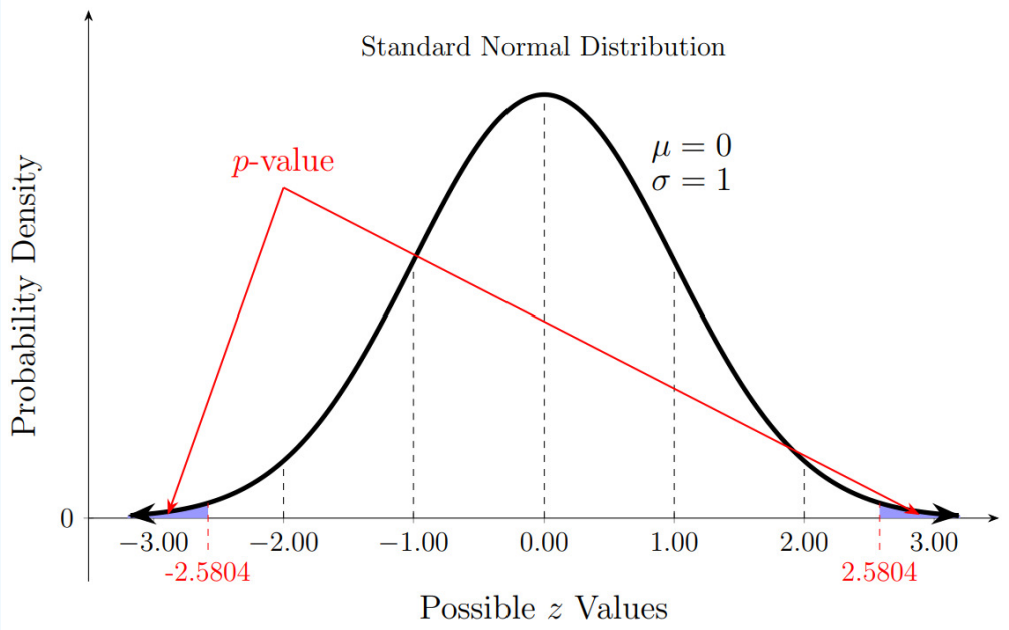


Figure 7.4.4 Standard normal distribution

$$p\text{-value} \approx 2 \cdot \text{NORM.S.DIST}(-2.5804, 1) \approx 0.0099$$

The p-value is just below the significance level of the test. We, therefore, have sufficient evidence to reject the null hypothesis in support of the notion that the proportion of females among newborn babies is not 50%.

7.5: Claims on Population Variances - Optional Material

Learning Objectives

- Conduct hypothesis testing on claims regarding population variance using the p-value method on one-tailed tests
- Introduce the critical value method
- Conduct hypothesis testing on claims regarding population variance using the critical value method on two-tailed tests

Section 7.5 Excel File: (contains all of the data sets for this section)

Review and Preview

Having developed hypothesis testing for claims on population means, paired variables, and proportions, we are aware that the process is supported by our understanding of the sampling distributions of particular sample statistics. This remains the case, when considering claims on population variance and standard deviation. Recall that the sample standard deviation is not an unbiased estimator of the population standard deviation but that the sample variance is an unbiased estimator of the population variance. Therefore, to test any claims on a population's standard deviation, we must first translate them into equivalent claims regarding the population's variance, test these new claims, and then translate the results back into the realm of standard deviation.

Once we formulate our hypotheses and collect our evidence, we assess the significance of the evidence using the p-value for one-tailed tests. Difficulties arise in determining the p-value when conducting a two-tailed test; they stem from determining what is equally extreme in the opposite direction when the distribution is not symmetric. We will address this difficulty in more detail later in the section and subsequently develop a different, yet common, approach to hypothesis testing. For the remainder of this introductory section, we will focus on the process for one-tailed tests because of its similarity to all that we have developed.

Recall that when the parent distribution is normal, we transformed the sampling distribution of sample variances into a χ^2-distribution with $n-1$ degrees of freedom to compute probabilities. We will need to utilize test statistics when testing claims on population variance. The test statistic is the value produced by mapping the evidence from a particular sample into the common distribution under the assumption that the null hypothesis is true. In assuming the null hypothesis is true, we will have some hypothesized value of the population variance, σ_0^2, leading to the following test statistic.

$$\chi^2_{n-1} = \frac{(n-1)}{\sigma_0^2} \cdot s^2$$

With the test statistic in hand, we compute the p-value and make a conclusion based on the comparison between the α value and the p-value. Let us begin testing claims on population variance and standard deviation.

Claims on Population Variance: One-Tailed Tests

In order to conduct hypothesis testing on claims regarding population variance, we will need to have a random sample taken from a normally distributed parent population. As with all hypothesis tests, checking that the requirements of the test are met is important! Let us consider an example situation together.

Many farmers spray their fields to prevent weeds and pests from negatively affecting their harvests. When spraying a field, it is important to get sufficient and even coverage. We need the average ratio of volume to area high enough to meet our needs and the standard deviation to be low enough to imply consistent application.

A company that manufactures sprayers conducted a test on a recently developed prototype to see if it met company standards regarding consistent, even coverage. The company will not produce a sprayer unless the standard deviation is less than a quarter of a gallon per acre. To test the consistency of the sprayer, the prototype sprayed three fields each containing 100 collection devices scattered sporadically throughout the field. When all was said and done, the 300 measurements averaged out to 15.3 gallons per acre with a standard deviation of 0.235 gallons per acre. They formulated the hypothesis test choosing a significance level of 0.10 and the following hypotheses.

$$H_0 : \sigma \geq 0.25 \text{ gallons per acre}$$
$$H_1 : \sigma < 0.25 \text{ gallons per acre}$$

This formulation of the hypotheses, however, is not the formulation that the company used in testing because the hypothesis testing needs to be done in the realm of variance, which yields the following set of hypotheses.

$$H_0 : \sigma^2 \geq 0.0625 \text{ gallons per acre}^2$$
$$H_1 : \sigma^2 < 0.0625 \text{ gallons per acre}^2$$

To conduct the hypothesis test, we need that the sample was randomly selected from a parent distribution that is normally distributed. Given the random placement of the 300 collection devices, the sample was randomly chosen. The company felt confident that the distribution was normally distributed based on past history, but they conducted a test on the sample data to see if it was reasonable based on the observed data (recall that such tests exist but are outside of the scope of this course). The test affirmed the reasonableness of the assumption that the parent distribution was normally distributed. So, the hypothesis test could be conducted. Note that the sample variance is 0.055225square gallons per square acre. We compute our test statistic and produce our visualization.

$$\chi^2_{299} = \frac{(300-1)}{0.0625} \cdot 0.055225 \approx 264.1964$$

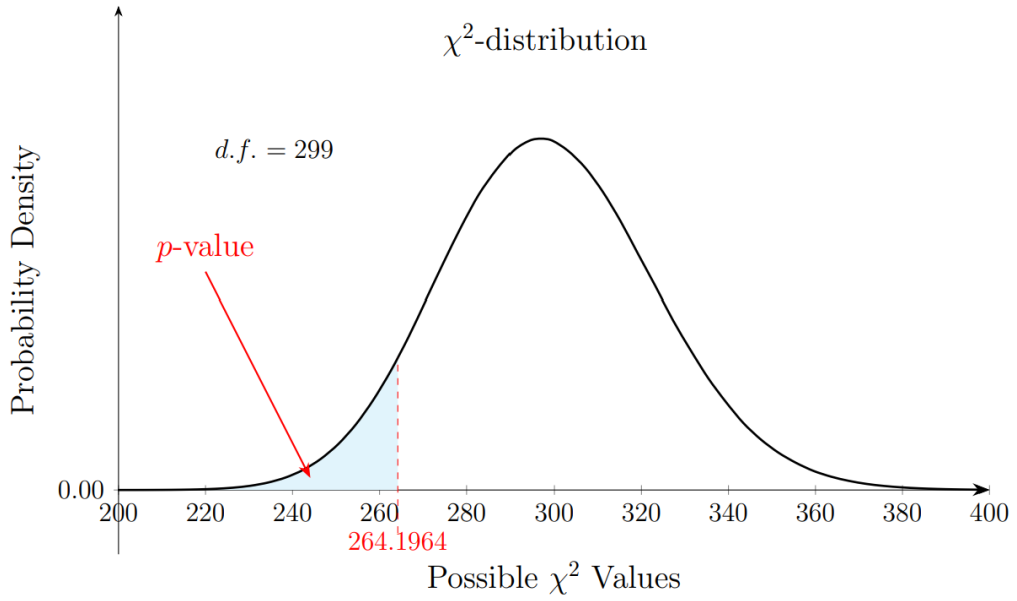


Figure 7.5.1: χ^2-distribution

We note that the χ^2-distribution appears to be symmetric as opposed to the asymmetrical appearance we have come to recognize. This is because the sample size is so large. The amount of skew present in χ^2-distributions decreases as the degrees of freedom increase. From our visualization we compute our p-value in order to conclude the hypothesis test.

$$p\text{-value} \approx \text{CHISQ.DIST}(264.1964, 299, 1) \approx 0.0728$$

Given that the level of significance for this test was 0.10, there is sufficient evidence to reject the null hypothesis. The company can begin to produce the first generation of this prototype sprayer.

? Text Exercise 7.5.1

An amateur game developer is designing a game with AI generated open worlds in hopes of building a game that is essentially endless. The developer does not, however, want the game to become monotonous and has tried to incorporate a great variability between worlds. One of the metrics the developer decided to use to test if the AI is producing enough variability is the distance the first significant encounter occurs from the starting position. The developer does not want the distance to be too long and does not want it to be too consistent. The developer designed the AI to produce worlds so that the average distance is about 550 game paces with a standard deviation of more than 170 game paces.

To make sure the AI was working properly, the developer randomly chose 50 game backers to play randomly chosen AI generated worlds in order to find the distances to the first significant encounter. The sample data was analyzed and was found to have an average distance of 497 game paces and standard deviation of 200 game paces. Test the hypothesis at the 0.05

significance level under the assumption that the distribution of the number of game paces to the first significant encounter is normally distributed.

Answer

We can conduct the hypothesis test because the sample was randomly selected and we were told to assume the parent distribution is normally distributed. The problem is framed within the context of standard deviation; so, we must translate the problem to variance. If the standard deviation is supposed to be more than 170 game paces, the variance would need to be more than $170^2 = 28,900$ square game paces. Since this game is just being developed and tested to see if it is working correctly, we do not want to assume that the population variance is greater than $28,900$square game paces. This helps us to set our hypotheses as follows.

$$H_0 : \sigma \leq 28,900 \text{ game paces}^2$$
$$H_1 : \sigma > 28,900 \text{ game paces}^2$$

We have a right-tailed test. We compute our test statistic using the sample variance and then produce our visualization to help compute the p-value.

$$\chi^2_{49} = \frac{(50-1)}{40,000} \cdot 28,900 \approx 67.8201$$

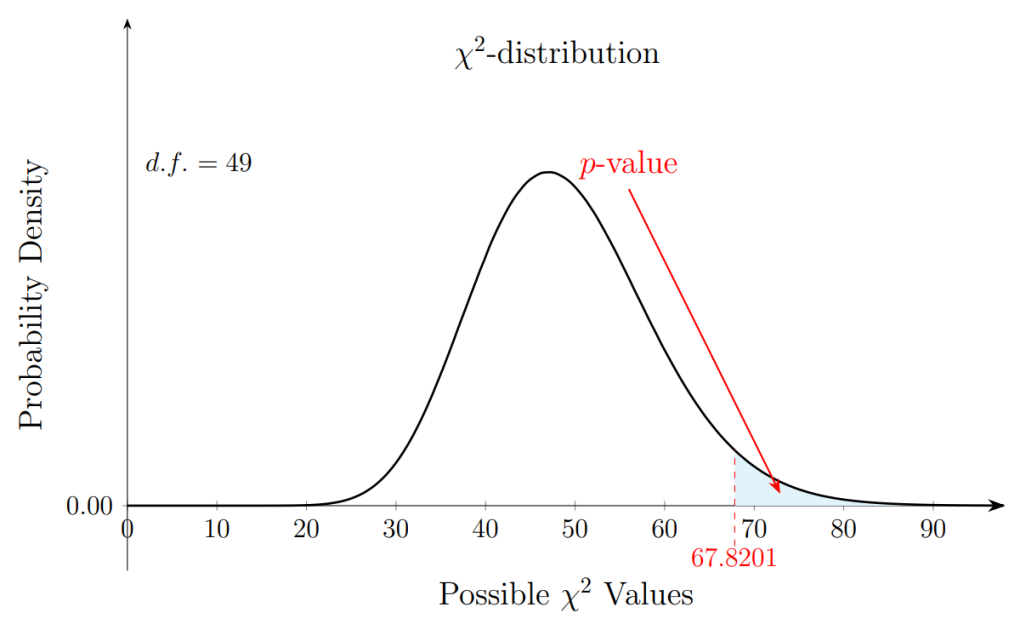


Figure $7.5.2$ χ^2-distribution

$$p\text{-value} \approx 1 - \text{CHISQ.DIST}(67.8201, 49, 1) \approx 0.0387$$

The p-value is smaller than the level of significance; therefore we reject the null hypothesis. This provides sufficient evidence for the developer to assert that the AI is working for the variation in the game. It looks like it may not be meeting specifications regarding the average distance though. That would require a test on means. An interested reader is encouraged to consider how to conduct such a test.

Claims on Population Variance: Two-Tailed Tests

As hinted in the Review and Preview section, we will take a separate approach to conducting two-tailed tests on population variances. This second technique can be applied to the hypothesis tests in general, but we leave such application to the reader. Let us examine why an issue arises with two-tailed tests on population variances. Recall two main ideas: the standard normal distribution and the t-distribution are symmetric about 0 (the expected value of each distribution) and the p-value is the probability of obtaining something at least as extreme as what was observed under the assumption that the null hypothesis is true. In the two-

tailed case, we needed to consider the value of a test statistic equally extreme as the test statistic computed from the observed sample statistic but in the opposite direction. We chose the value that was the same distance away from the mean just with the opposite sign. Given the symmetry of the previous distributions, three facts about the two values coincide: equidistant from the mean, equal probability in the tails, and the heights of the density function at those values match. As it turns out, these serve as three different possibilities for determining the value that would be equally extreme just in the opposite direction. Since the χ^2-distribution is not symmetric, these three facts do not coincide in the χ^2-distribution. Arguments can be made for the legitimacy of each possible definition; we leave such discussion for more advanced studies, and instead introduce a method common to many textbooks that can be applied just as easily in this context as in the other contexts considered thus far in the book.

Critical Value Method

The p-value and critical value methods share much in common: the requirements to conduct the hypothesis test, the designation of an α value, and the computation of a test statistic under the assumption that the null hypothesis is true to name a few. The primary difference lies in how to assess the significance of the collected evidence. In the p-value method, we compare the probability of getting something at least as extreme as what was observed to the α value. If the p-value is less than the α value, we have sufficient evidence to reject the null hypothesis. In the critical value method, we determine, based on the α value, what values of the test statistic constitute significant evidence. We segment the distribution of the test statistics into regions based on whether or not we will reject or fail to reject the null hypothesis if the computed test statistic falls in them or not. The regions where we would reject the null hypothesis are called rejection regions. The boundary points of these regions are called critical values, hence the name of the method. We must address how to identify these regions and set their boundaries.

If we are conducting a right-tailed test, we are looking for evidence against the null hypothesis by looking for test statistics far to the right of the expected value. If we are conducting a left-tailed test, we are looking for evidence against the null hypothesis by looking for test statistics far to the left of the expected value. If we are conducting two-tailed tests, we are looking for evidence against the null hypothesis by looking for a test statistic differing from the expected value in either direction. From these thoughts, we identify our rejection regions. If we have a right-tailed test, our rejection region lies in the right tail. If we have a left-tailed test, our rejection region lies in the left tail. And, similarly, if we have a two-tailed test, our region region has two components: both the left and right tails.

But how far along the tails must the computed test statistic be in order for us to fall in the rejection region? It depends on the α value. The smaller the α is the farther along the tail we need the computed test statistic to fall. Recall that we can understand the α value as the probability of making a type I error given that the null hypothesis is actually true. Once we have selected a particular α value for a test, that value represents the expected rate of making a type I error if the null hypothesis is true and we repeatedly collect random samples to test the hypothesis. We thus determine the size of our rejection region by setting the probability of a test statistic falling in the region to be the α value. In one-tailed tests, the entire area naturally falls in one tail, but with two-tailed tests, the area must be split between the two tails; each having an area of $\frac{\alpha}{2}$.

We have the following formulation of the critical value method.

1. Use natural observation, previous experimental results, or the claims of others to formulate a hypothesis that warrants testing.
2. Identify a competing hypothesis. Set the null and alternative hypotheses.
3. Set the α value for this particular hypothesis test
4. Determine the methodology of collecting evidence against the null hypothesis and determine what constitutes sufficient evidence by setting the level of significance. Make sure the design meets the requirements of the tests intended to be conducted.
5. Conduct the experiment and collect the evidence.
6. Compute the test statistic.
7. Use the hypotheses to determine whether a test is a left-tailed, right-tailed, or two-tailed test. Note that the directions match with sign in the alternative hypothesis.
8. Determine the rejection region of the appropriate distribution of the test statistics based on the hypothesis.
9. Determine if the test statistic falls within the rejection region. If so, reject the null hypothesis. If the test statistic falls on the boundary or outside of the rejection region, fail to reject the null hypothesis.

When using the same α value, the critical value method will produce the same conclusions to hypothesis tests as the p-value method when conducting one-tailed tests on claims regarding means, proportions, and variances and when conducted two-tailed tests on claims regarding means and proportions. The p-value method is the most prevalent method in part because simply relaying

the p-value allows the readers to assess the strength of the evidence and to apply their personal thresholds without additional work which is not realistic with the critical value method.

Text Exercise 7.5.2

When conducting hypothesis tests regarding claims on population variance, there are three forms that the hypotheses can have and there are two conclusions that can be drawn from each form. This yields six total possibilities. The following pictures visualize the implementation of the critical value method for claims on population variance with a random sample of size n taken from a normally distributed parent population. Note that the critical values are denoted using similar notation as the critical values used in constructing confidence intervals for population variance, the rejection region is colored a light red, and the test statistic is denoted χ^2_{n-1}. For each picture, deduce the formulation of the hypotheses and determine the conclusion of the test.

1.

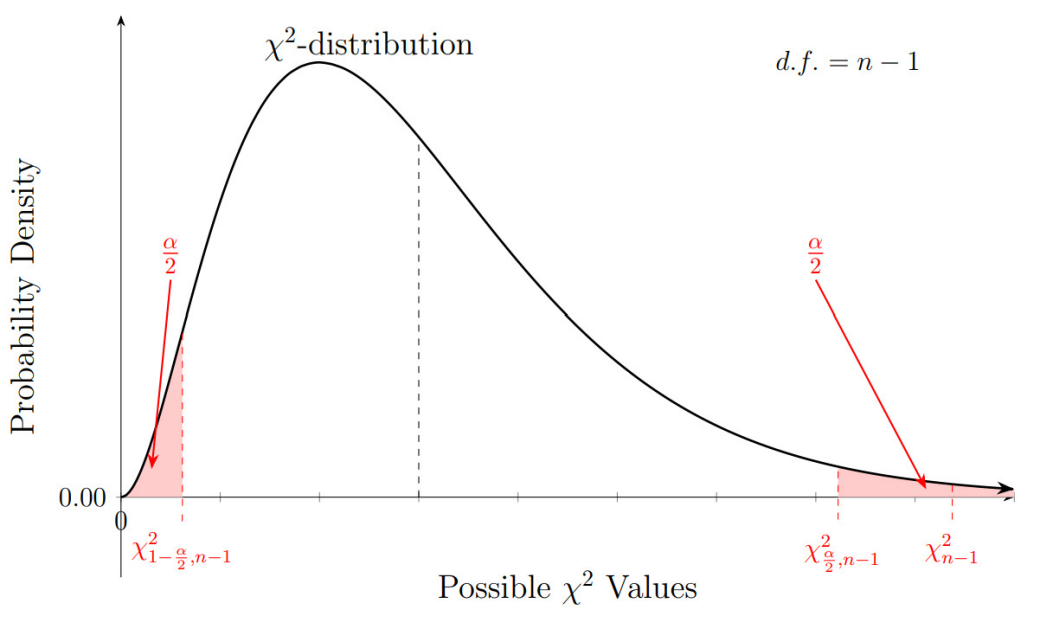


Figure 7.5.3: χ^2-distribution

Answer

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

Since the critical value is in the shaded rejection region, we have sufficient evidence to reject the null hypothesis.

2.

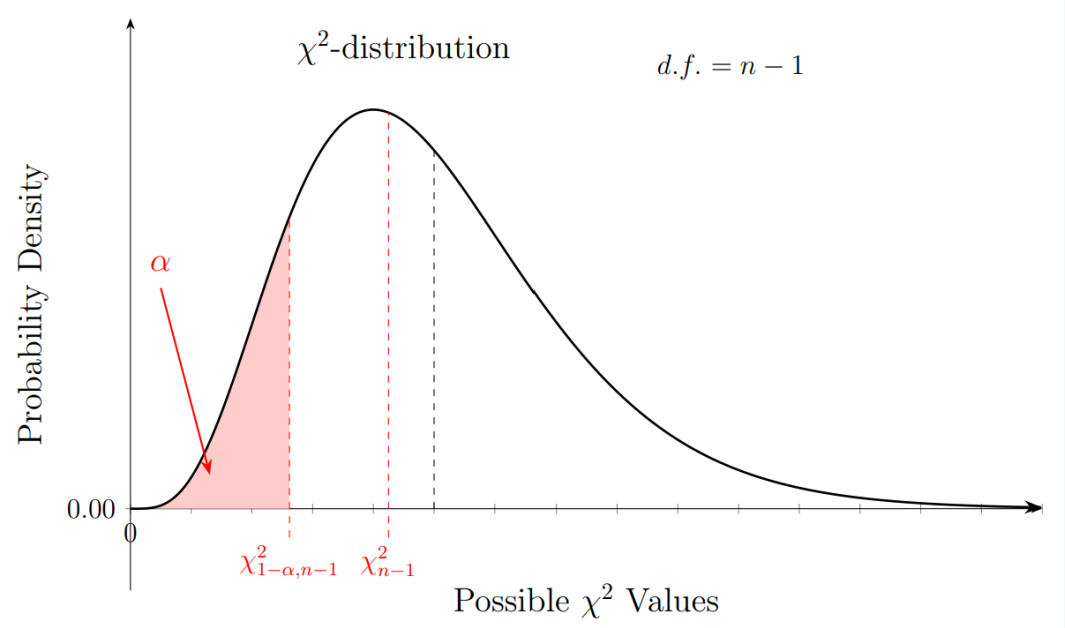


Figure 7.5.4: χ^2-distribution

Answer

$$H_0 : \sigma^2 \geq \sigma_0^2$$
$$H_1 : \sigma^2 < \sigma_0^2$$

Since the critical value is not in the shaded rejection region, we do not have sufficient evidence to reject the null hypothesis.

3.

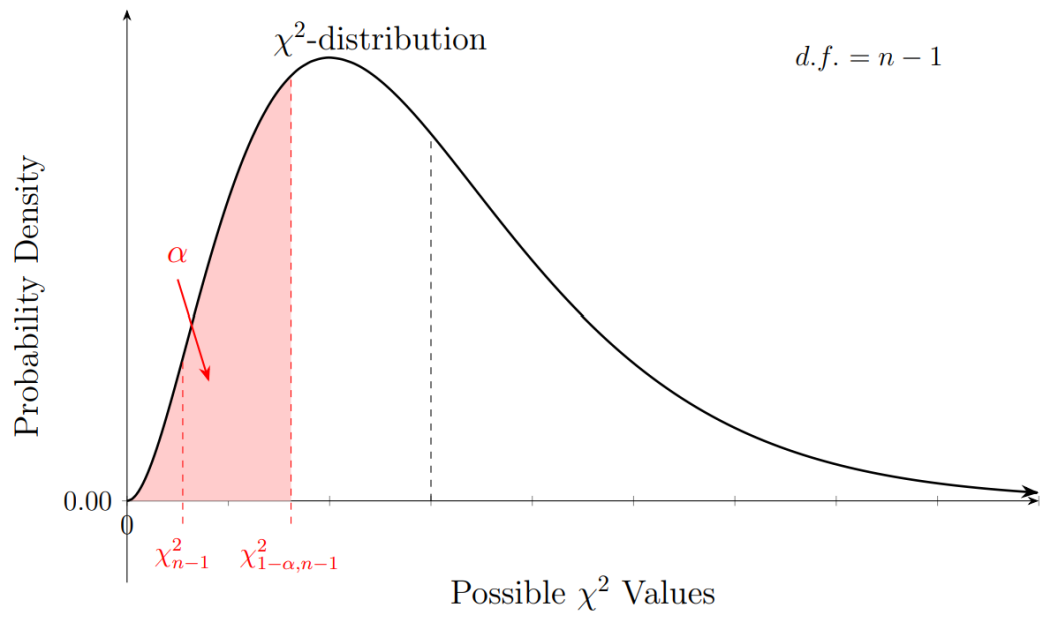


Figure 7.5.5: χ^2-distribution

Answer

$$H_0 : \sigma^2 \geq \sigma_0^2$$
$$H_1 : \sigma^2 < \sigma_0^2$$

Since the critical value is in the shaded rejection region, we have sufficient evidence to reject the null hypothesis.

4.

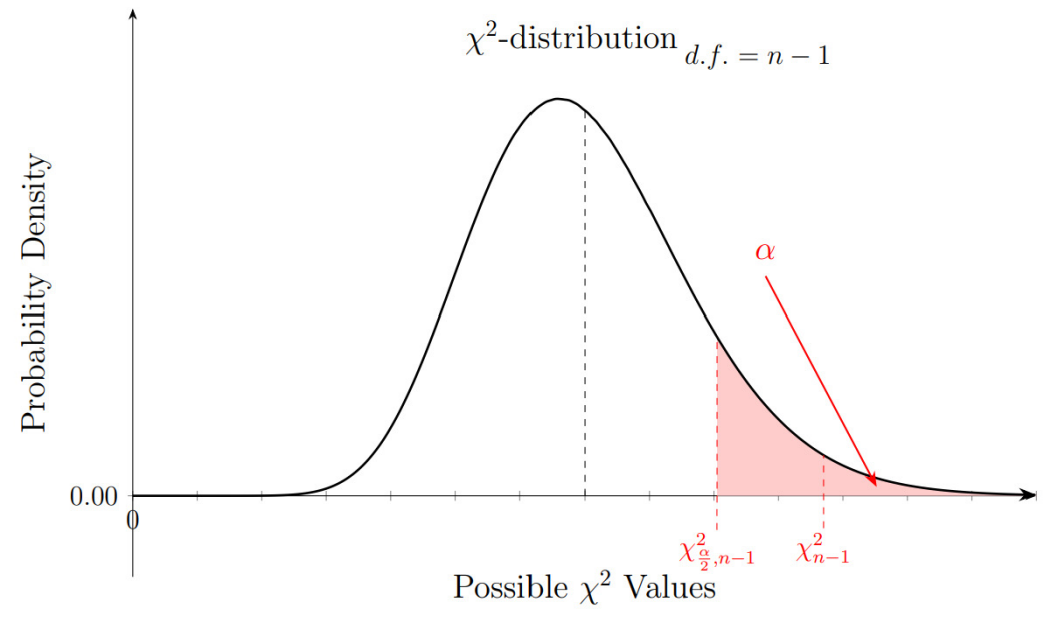


Figure 7.5.6: χ^2-distribution

Answer

$$H_0 : \sigma^2 \leq \sigma_0^2$$
$$H_1 : \sigma^2 > \sigma_0^2$$

Since the critical value is in the shaded rejection region, we have sufficient evidence to reject the null hypothesis.

5.

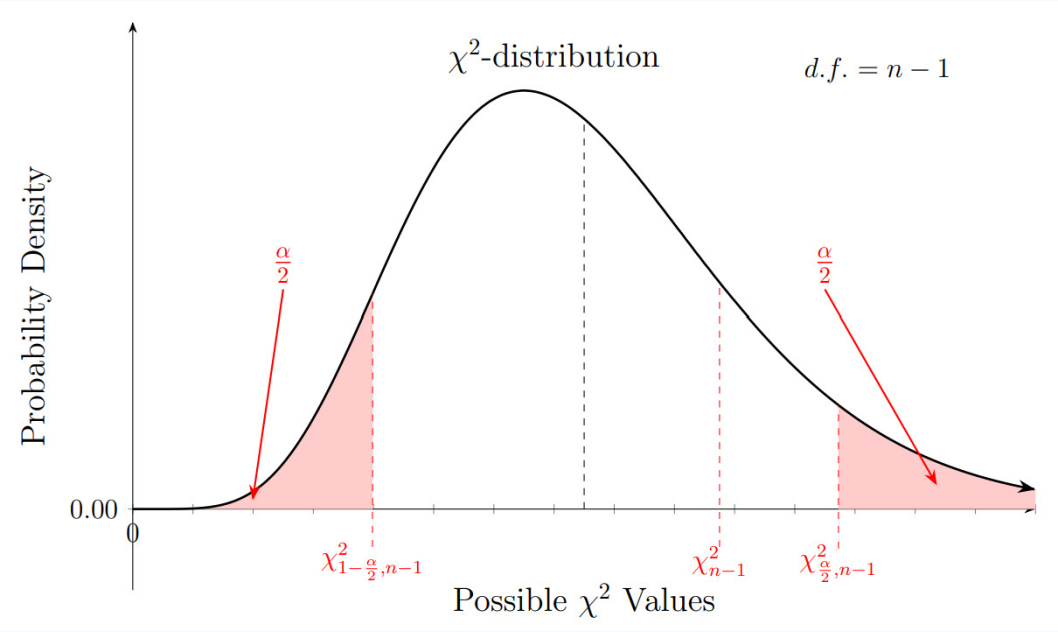


Figure 7.5.7: χ^2-distribution

Answer

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

Since the critical value is not in the shaded rejection region, we do not have sufficient evidence to reject the null hypothesis.

6.

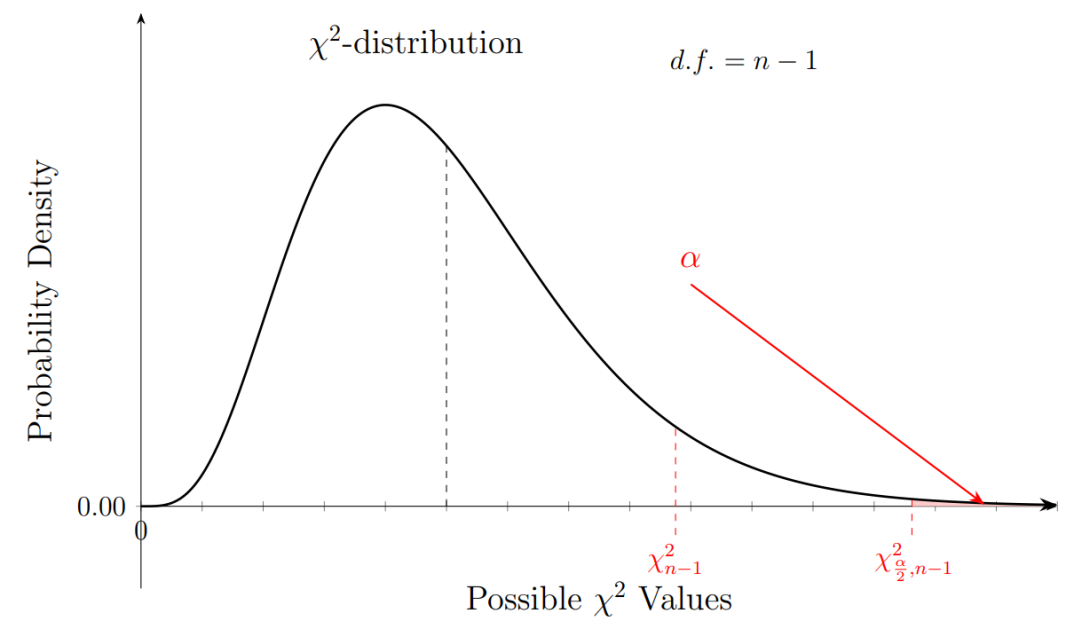


Figure 7.5.8: χ^2-distribution

Answer

$$H_0 : \sigma^2 \leq \sigma_0^2$$
$$H_1 : \sigma^2 > \sigma_0^2$$

Since the critical value is not in the shaded rejection region, we do not have sufficient evidence to reject the null hypothesis.

Text Exercise 7.5.3

In Text Exercise 5.2.3, we claimed that the heights of adult females followed a normal distribution with an average height of 64 inches and a standard deviation of 2.5 inches. A researcher thinks that the variation of adult female heights changes with time due to a combination of genetics, nutrition, and lifestyle. The researcher decides to test this claim at a level of significance of 0.01 by randomly sampling 15 adult females. Their heights are reported below. Conduct the test using the critical value method.

$$59, 59, 61, 62, 63, 63, 64, 64, 65, 66, 68, 69, 69, 69, 70$$

Answer

The heights of adult females are known to be normally distributed and the sample was randomly selected. We can, therefore, conduct the hypothesis test. Since the researcher is interested in any difference in the variability, we will have a two-tailed test. We do not want to assume that the researcher is correct without evidence. We settle on the following hypotheses.

$$H_0 : \sigma^2 = 6.25 \text{ inches}^2$$
$$H_1 : \sigma^2 \neq 6.25 \text{ inches}^2$$

We now compute the sample variance from the collected data and arrive at $s^2 \approx 13.4952$ square inches. We compute our test statistic.

$$\chi^2_{14} \approx \frac{15-1}{6.25} \cdot 13.4962 \approx 30.2293$$

In order to compute the critical values, we recall that the degrees of freedom are $n-1$ and that we must split the α equally between the two tails. This means that only $\frac{0.01}{2} = 0.005$ will be in each tail. We compute the critical values using technology.

$$\chi^2_{0.005,14} = \text{CHISQ.INV}(0.005, 14) \approx 4.0747$$

$$\chi^2_{0.995,14} = \text{CHISQ.INV}(0.995, 14) \approx 31.3194$$

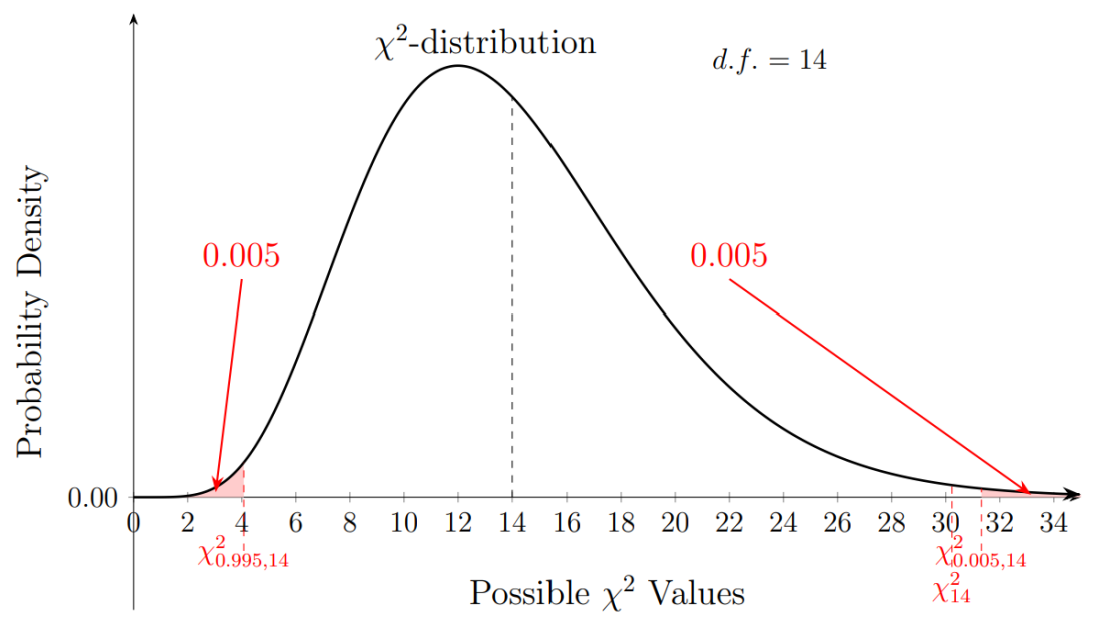


Figure 7.5.9 χ^2-distribution

The test statistic is greater than the smaller critical value while being smaller than the larger critical value. The test statistic, therefore, falls in the fail to reject region of the distribution of test statistics. We conclude that there is not sufficient evidence to reject the null hypothesis. We cannot affirm the researchers' claims that the variability present in adult female heights is different than it once was with a standard deviation of 2.5 inches.

CHAPTER OVERVIEW

8: Linear Correlation and Regression

8.1: Introduction to Bivariate Quantitative Data

Learning Objectives

- Define multivariate data
- Introduce, differentiate, and identify associations and correlations
- Construct and utilize scatter plots to analyze bivariate data
- Distinguish between a linear and a nonlinear relationship
- Explore the differences between causation and correlation

Section 8.1 Excel File: (contains all of the data sets for this section)

Review and Preview

When we observe the world around us, there are a multitude of questions that could be asked about any single object, person, or event. So, when we study a population, it is possible to have many varied interests in the population or each member of the population. Take, for example, the assessment of the general health of an individual by a doctor. When we go to a healthcare provider for an assessment of general health, the doctor considers more than just our height. Multiple factors, like age, sex, height, weight, cholesterol, and glucose, to name a few, are considered. To get an accurate understanding of our general health, multiple variables must be considered together. We collect multivariate data when we are interested in a set of variables from each individual being studied. As this is just an introductory text, we will limit our considerations to bivariate quantitative data, meaning that we only consider analyses with only two quantitative variables of interest.

Bivariate Data: Types of Association and Models

Consider the ages at which married couples gave their wedding vows. For each married couple, we are interested in both the age of the bride and the age of groom. As such, we are considering bivariate data. We sampled 15 different married couples and tabulated the data below.

Table 8.1.1: Age of bride and groom on wedding day

Married Couple	Groom's Age (years)	Bride's Age (years)
1	20	21
2	26	20
3	32	34
4	30	30
5	21	22
6	29	28
7	26	25
8	34	34
9	29	28
10	55	50
11	30	26
12	43	39
13	30	29
14	24	22
15	20	19

Even with just 15 married couples, the data is difficult to digest and summarize. In taking the time to compare the ages of the bride and groom for each married couple, we come to the inclination that it is fairly common for the ages to be somewhat close together. This inclination aligns with the intuition built from previous experience. As such, we expect that there is a relationship between the age of the groom and the age of the bride. When the bride is young, we expect a young groom. When the groom is old, we expect an old bride. Given this expected pattern in the ages, we describe the association as positive. If we consider the ages of the bride and groom as quantitative variables, then as one variable increases, we expect the other variable to increase as well. Equivalently, if one variable decreases, we expect the other variable to decrease. As one variable changes in amount, the other variable changes in the same direction. As such, positive associations between variables are often referred to as direct relationships or positively correlated.

We built an intuition for this relationship with our previous experiences and by comparing the data line-by-line. Bivariate data will not always center around topics so commonplace and with so few observations. We will need to develop methodologies for facilitating such comparisons both visually and analytically. A common way to visualize bivariate quantitative data is by constructing a scatter plot. We are interested in establishing the relationship between the age of the bride and the age of the groom on their wedding day. The first column of the data simply labels the married couples and does not provide any information regarding the desired relationship. We focus on the second and third columns of our tabulated data. We can treat each married couple as a coordinate pair (Groom's Age, Bridge's Age) and then plot the 15 points on a coordinate plane to obtain a visualization of the relationship between the quantitative variables age of groom and age of bride. We have done so below.

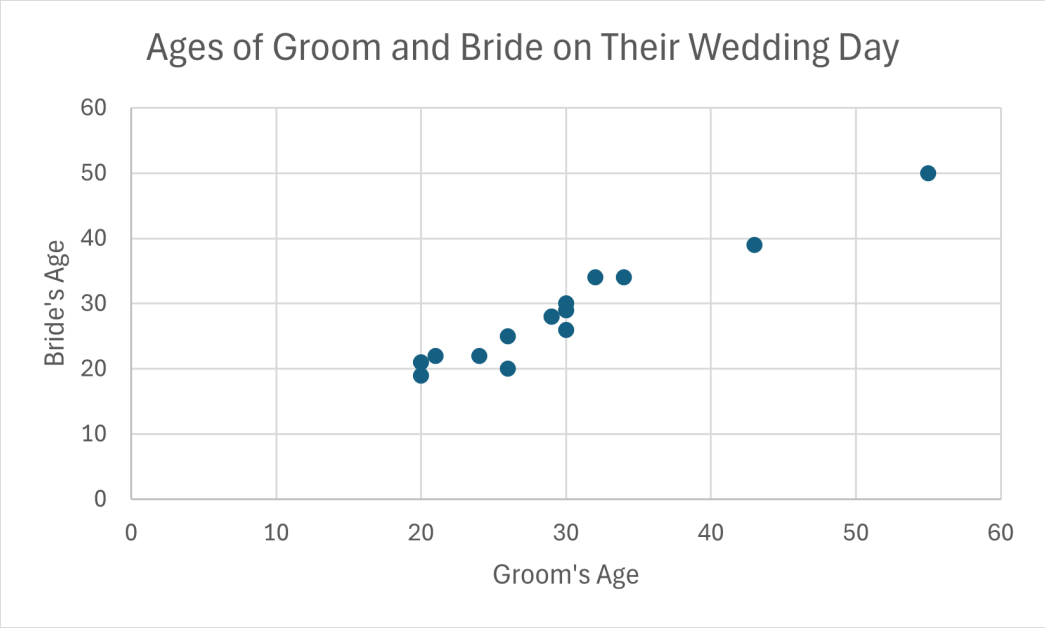


Figure 8.1.1: Scatter plot of groom's age vs bride's age on wedding day

The scatter plot quickly and easily confirms our initial intuitions. There is a relationship between the age of the groom and the age of the bride on their wedding day. As one variable increases, so does the other. Recall that our initial thoughts were that the ages were fairly close together and, as a result, we concluded that as one increased so would the other. We need to take a closer look to see how closely they are related. If the ages of the groom and bride are close together, we would expect that the data points would fall close to the line where the age of the groom equals the age of the bride. We have plotted such a line on the scatter plot below.

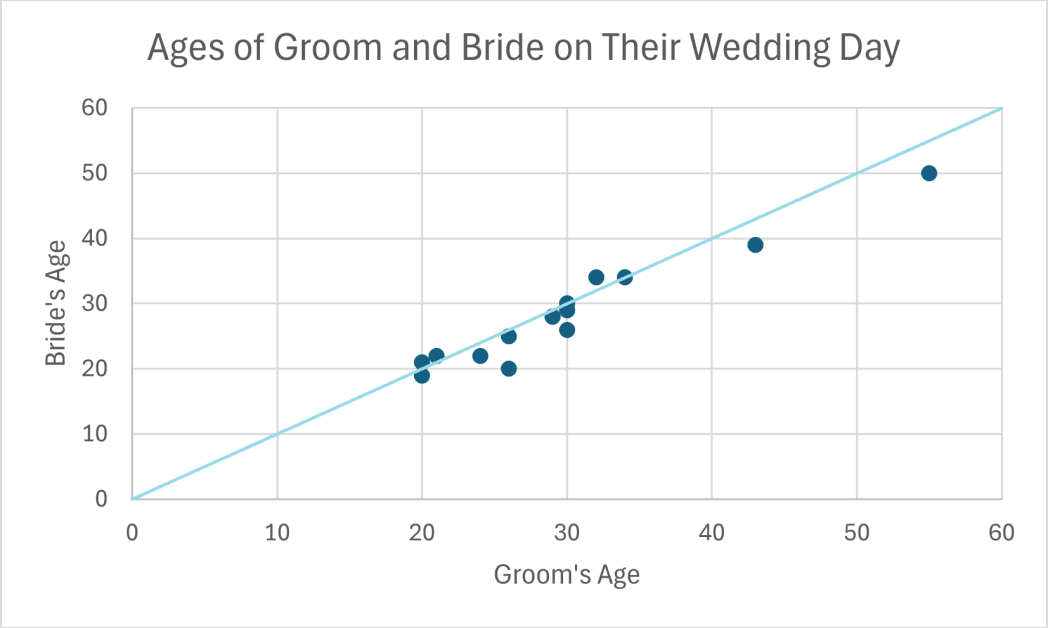


Figure 8.1.2: Scatter plot with the line $y = x$ of groom's age vs bride's age on wedding day

The line fits the data fairly well and gives credence to the idea that the relationship between the ages of the bride and groom can be modeled using a linear function. Recall that a linear function (straight line), is characterized by having a rate of change, a constant slope. The slope is the ratio of the vertical change to the horizontal change (rise over run). Since this is constant, we expect that the change in one variable is proportional to the change in the other variable meaning $\delta y = m\delta x$, where m is the slope of the line, regardless of the particular values of the variables. This seems to be true of the scatter plot at hand, but perhaps there is another line, with a different slope or y-intercept, that represents the data better. Consider the following scatter plot with the additional linear function represented using the blue dotted line.

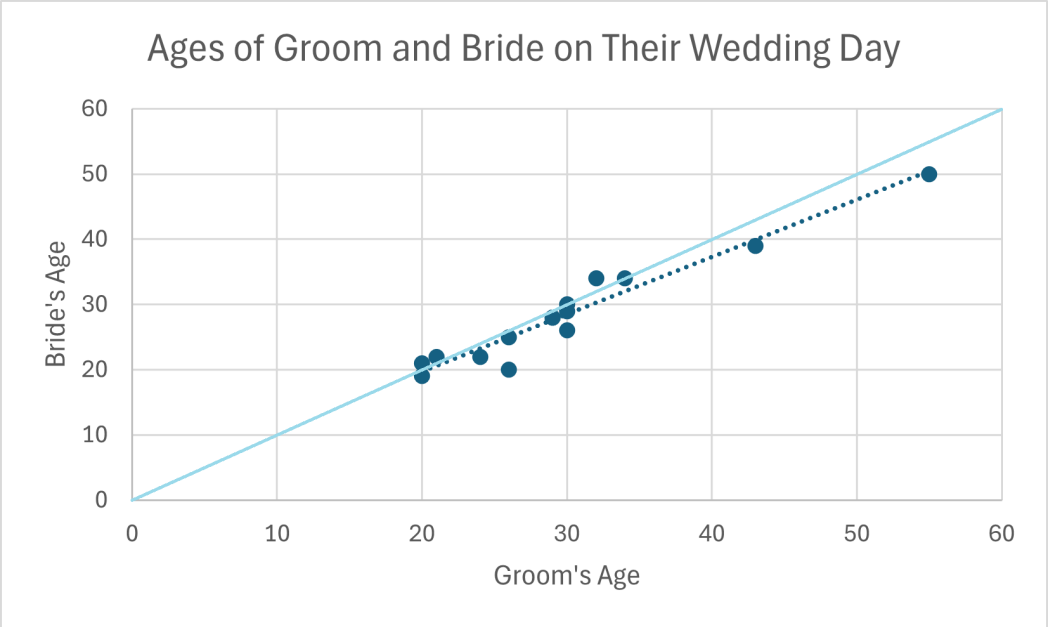


Figure 8.1.3: Scatter plot with two lines of groom's age vs bride's age on wedding day

Both of these lines appear to fit the data fairly well. We will eventually ask the question of how to decide which is better. As of right now, we conclude that when considering married couples, the two quantitative variables age of bride and age of groom appear to have a direct relationship and the relationship is likely to be modeled using a linear function.

Text Exercise 8.1.1

1. When we were considering the bivariate data related to the ages of the bride and groom on their wedding day, we asserted that the two quantitative variables displayed a positive correlation. It is possible to have a negative correlation (also referred to as a negative association or inverse relationship) between two quantitative variables. A negative correlation occurs when, as one variable increases, the other variable decreases, or equivalently: as one variable decreases, the other increases. Remaining within the context of marriage, identify bivariate data that would likely display a negative correlation. Explain your reasoning.

Answer

Given such a broad topic, the answers can vary quite tremendously. To check your solution, ensure that, for each member of a sample, two quantitative measurements are taken. This makes the quantitative data bivariate. Consider the relationship between the values of the two measurements. Will large values of one variable correspond to small values in the other? Does one decrease as the other increases or vice-versa?

A rather simple example relates the number of children in the household with the amount of one-on-one time a husband and wife get to spend with each other. As the number of children in the household increases, the responsibilities of the husband and wife as parents occupy a greater proportion of their time. Most couples realize the need to continually spend time together; so, we would not expect the amount of one-on-one time to diminish to 0, but nonetheless, real and felt decreases in one-on-one time is expected as the family grows in size.

2. Remaining in the context of marriage, give an example of bivariate data in which there is little to no correlation. That would mean if one variable is large, the other may be either small or large, and as one variable is small, the other may be small or large.

Answer

Consider the average height of the couple along with their annual income. In principle, these two variables have nothing to do with each other. While we could imagine reasons why these two quantities may correlate, there are many other factors, such as genetics, which determine height. We could reasonably expect to see short couples with low income, short couples with high income, tall couples with low income, and tall couples with high income.

Just like before, there are innumerably many possible answers to this question. If you picked two quantities where knowing what one is does not inform what the other will be, then your example likely works.

So far, we have described the concept of association within bivariate quantitative data as whether or not there is a relationship between the two variables. If there is an association, we are interested in what generally happens or is expected to happen to the value of one variable as the other variable is changed. If the directions of the changes match, the association is said to be positively correlated. If the directions of the changes are opposite, the association is said to be negatively correlate. It is possible to have relationships between two variables that have positive associations on certain intervals and negative associations on others. In such a case there is an association because there is a relationship, but there is no correlation because the relationship between the variables cannot be simply described as increasing or decreasing. Such relationships, however, would not be modeled well by a single linear function and, therefore, fall out of the scope of this course.

Text Exercise 8.1.2

The following questions center around the bivariate data related to diamonds, the gemstones that point to unwavering and lifelong love, according to popular culture. The quality of a diamond depends on various factors: the cut, color, clarity, and weight. The general shape or cut of the diamond plays an important role, but when the quality of the cut is referenced, the focus is on the proportions of the cut diamond as they relate to reflecting light back through the diamond. The color points to the general hue of the diamond; while, the clarity points to the presence of internal or surface defects on the diamond. The weight of the diamond is typically measured in carats. The prices of various round diamonds with super ideal cuts, flawless clarity, and icy white hue were observed along with their weights from several top national retailers (Brilliant Earth and Blue Nile). The diamonds were then ordered to produce the following table.

Table 8.1.2: Weight (carats) and price ($) of diamonds

Diamond	Weight (carats)	Price (US dollars)
1	0.37	1, 160
2	0.52	2, 430
3	0.63	3, 860
4	0.63	3, 430
5	0.68	3, 900
6	0.7	4, 050
7	0.77	5, 020
8	0.83	6, 830
9	0.95	9, 560
10	1.18	12, 830
11	1.24	13, 610
12	1.3	14, 830
13	1.3	14, 890
14	1.37	15, 990
15	1.43	18, 890
16	1.61	24, 000
17	1.67	23, 870
18	2	39, 300
19	2.02	38, 580
20	2.22	54, 360
21	2.23	63, 820
22	2.34	60, 160
23	2.39	49, 130
24	2.52	65, 600
25	2.56	62, 130
26	3.5	144, 120
27	3.56	192, 710
28	3.88	157, 280
29	4.42	236, 360
30	5.02	322, 260
31	5.06	374, 870
32	5.8	359, 600
33	6.04	398, 190
34	7.13	543, 230

Diamond	Weight (carats)	Price (US dollars)
35	7.32	530, 320

1. What sort of association do you expect with this bivariate data? Explain.

Answer

As diamonds increase in size and weight, the rarity of the diamonds increases. And as rarity increases, the price increases. We expect a positive correlation.

2. Construct a scatter plot of this bivariate data to check your intuition from the previous part of this exercise. Be sure to label the graph and both axes.

Answer

Using the data tabulated in the provide Excel file and under the guidance of the Excel guide, we construct the following scatter plot.

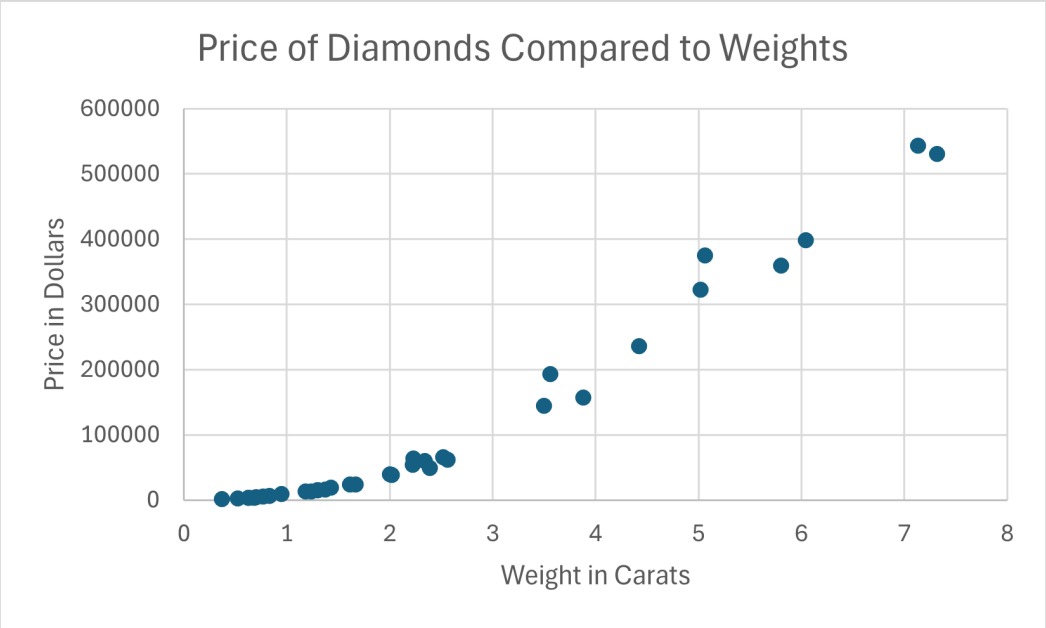


Figure 8.1.4 Scatter plot of weight (carats) vs price ($) of diamonds

As predicted, we see that as the weight of diamond increases, the price of the diamond also increases. We have confirmed that the correlation is positive.

3. If we were trying to model the association of this bivariate data with a function, would a linear function fit the data well?

Answer

In looking at the scatter plot, it appears that the rate at which the price is increasing as the weight increases is not constant. The rate of change when the weight is between 0 and 2 carats is perhaps fairly steady but is different from the rate of change when the weight is between 3 and 8 carats. Consider the slopes of the two lines drawn below.

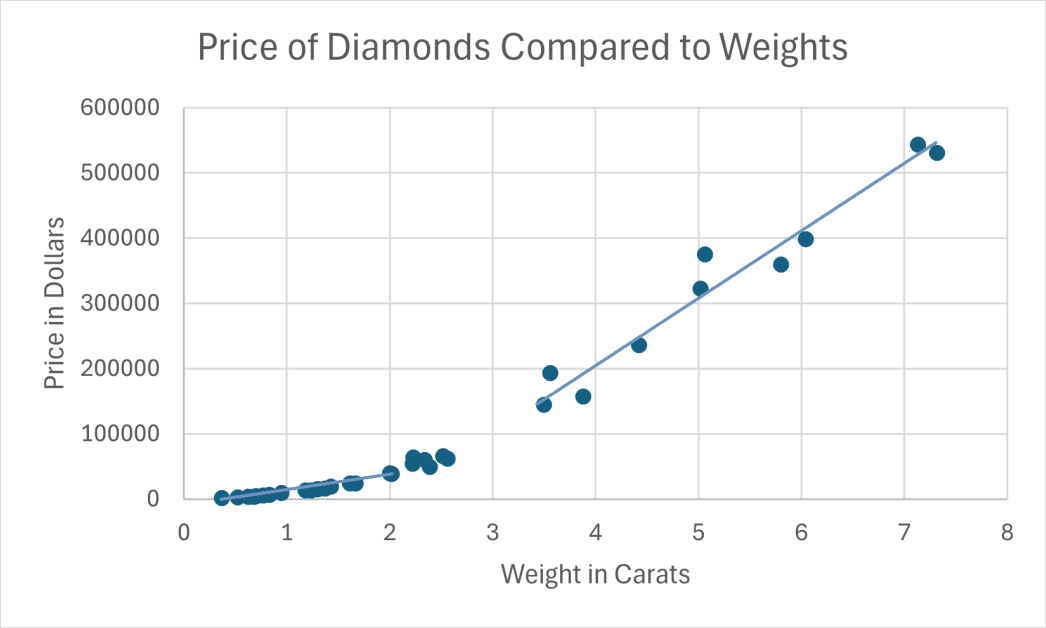


Figure 8.1.5 Scatter plot of weight (carats) vs price ($) of diamonds comparing slope

The slope of the line of the left is positive but significantly less steep than the slope of the line on the right. This leads us to conclude that a linear function is probably not the best model for the bivariate data.

We will limit our discussion in the text to using linear functions, but there are other types of functions that can be used as well. We modeled this same data using a power function to produce the following scatter plot and model below. Interested readers are encouraged seek more advanced statistical texts to address such content.

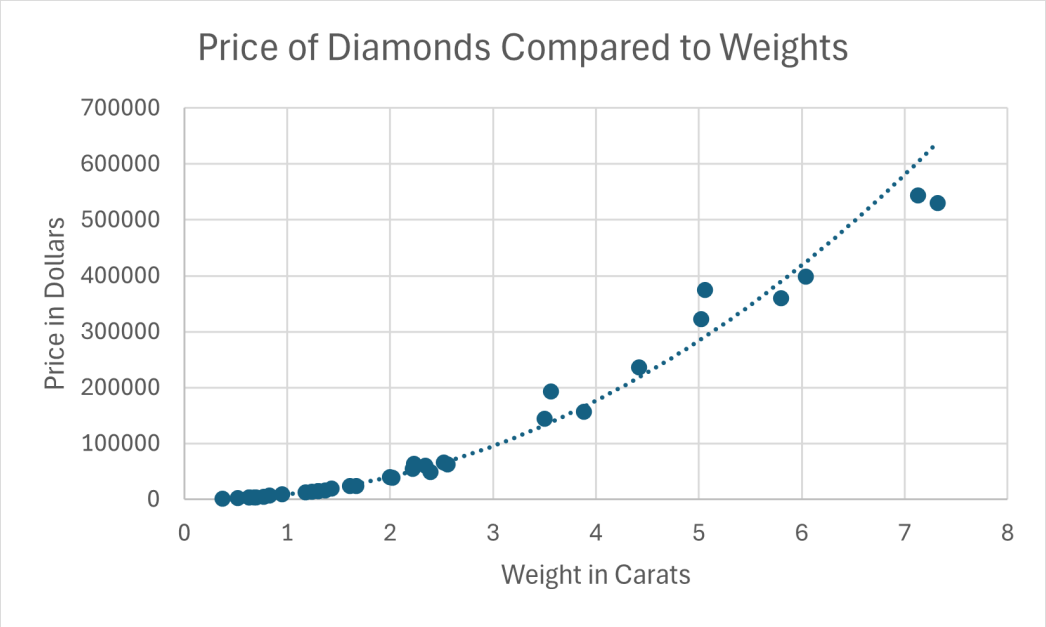


Figure 8.1.6 Scatter plot of weight (carats) vs price ($) of diamonds power function

Bivariate Data: Strength of Association

There are many instances when the relationship is not as clear as we have seen so far; the association is not as strong. Indeed there are cases, where there is no association. Consider the following scatter plot which relays the measurements the age of the bride at

the time of marriage with the number of children the couple has over the course of their marriage from a random sample of married couples.

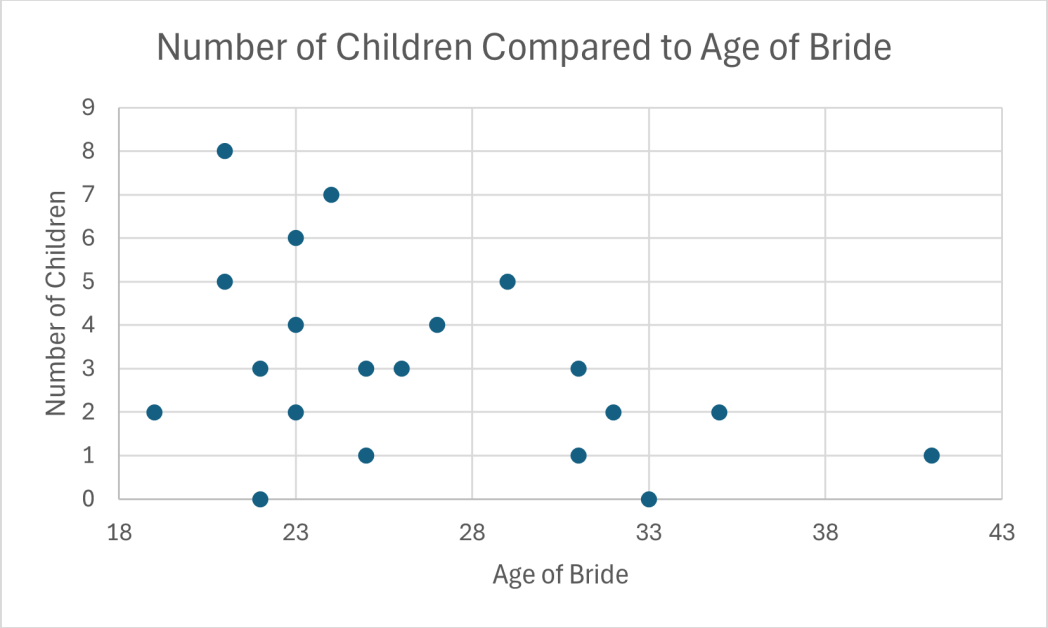


Figure 8.1.7: Scatter plot of age of bride vs number of children

Data in Tabulated Form

Table 8.1.3 Table of age of bride vs number of children

Married Couple	Bride's Age (years)	Number of Children
1	19	2
2	21	8
3	21	5
4	22	0
5	22	3
6	23	6
7	23	4
8	23	2
9	24	7
10	25	1
11	25	3
12	26	3
13	27	4
14	29	5
15	31	3
16	31	1
17	32	2

Married Couple	Bride's Age (years)	Number of Children
18	33	0
19	35	2
20	41	1

The scatter plot shows a general decline in the number of children as the age of the bride increases. This has a quite natural explanation from a basic understanding of human biology. This does not, however, explain the fact that our data looks as it does. Up until now, the coordinate pairs on our scatter plot looked somewhat like closely packed paths through the coordinate plane. Now our scatter plot looks similar to a shaded triangle. There are couples with 0, 1, and 2 children all throughout the presented age range. There are many factors that contribute to the number of children born through a marriage. There is a negative association between the age of the bride and the number of children, but the strength of the association is not as strong as the other examples we have seen due to a number of other factors. The closer the points are clustered to form a tightly packed path, the stronger the relationship; the less densely the data is packed along a path, the weaker the relationship. If there is no path, there is no relationship. We would see a similar sort of loss of strength if in our search for diamond prices and weights we did not first narrow our search to diamonds of the same cut, color, and clarity.

Text Exercise 8.1.3

For each scatter plot, indicate the type and strength of the association between the two variables and if a linear function would model the relationship well.

1.

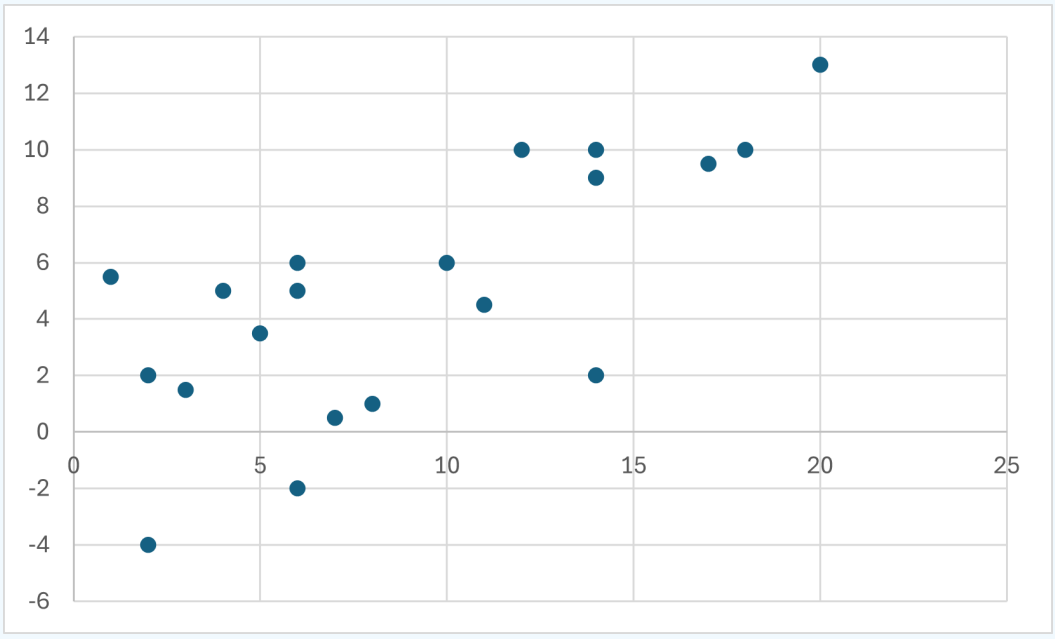


Figure 8.1.8: Scatter plot

Answer

The scatter plot indicates a positive association between the two variables, but the points on the scatter plot are not tightly packed together vertically. This leads us to say that the association is not as strong as some of the other associations seen in this section. It does look like a linear function could be used to model the data.

2.

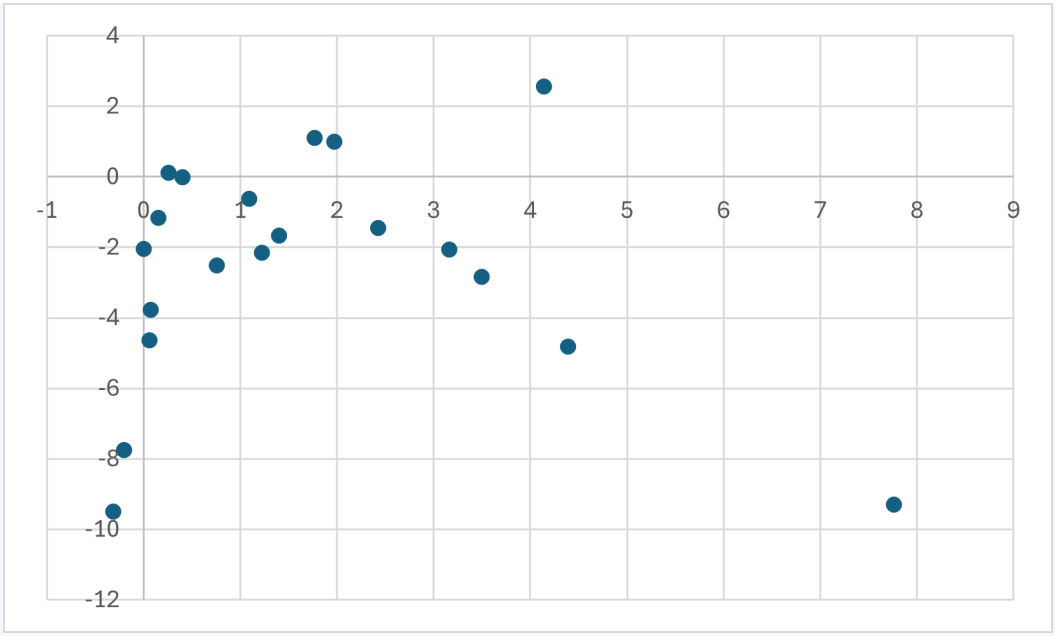


Figure 8.1.9: Scatter plot

Answer

The scatter plot indicates that the variables show a positive association when the horizontal variable is less than about 2 but a negative association after that. Given that the relationship does not appear to be a simply positive or negative association, we would say there is no correlation. The scatter plot does indicate that an association between the variables is present, possibly to a stronger degree than the previous part of this text exercise. Modeling this data would be best be done with a function that is increasing and then decreasing. As such, a linear function would not model the data well.

3.

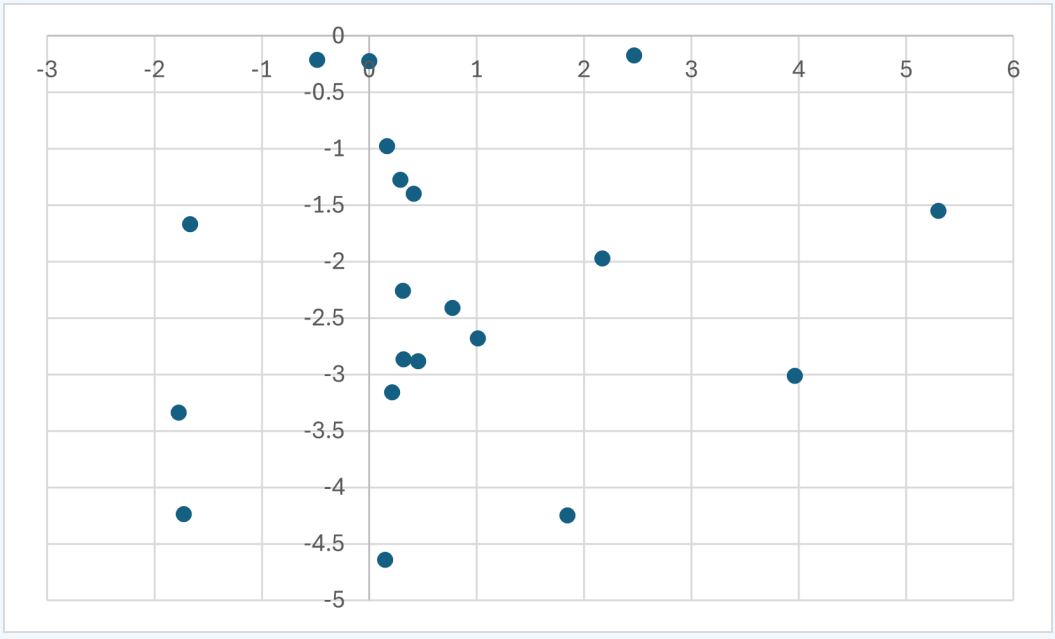


Figure 8.1.10: Scatter plot

Answer

This scatter plot does not seem to admit to any particular association. The data seems scattered fairly randomly over the coordinate plane. It is possible to envision a steep line with negative slope through the origin fitting the data okay, but it

leaves a lot of points farther away from the line, and it seems just as easy to envision a shallow line with positive slope through $(0, -2.75)$ fitting the data okay, but again leaving a lot of points far away from the model. As such, we say that there is likely no association, or at best, a very weak association between the variables. When there is no association or a very weak association, no function will serve as a suitable model.

4.

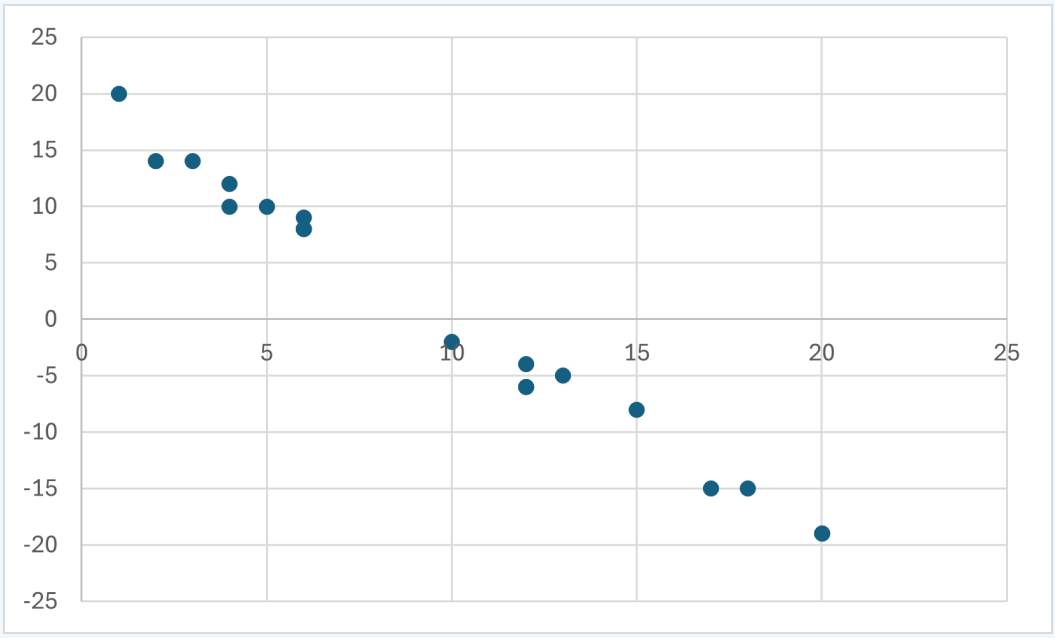


Figure 8.1.11: Scatter plot

Answer

This scatter plot indicates a negative correlation fairly clearly. The association appears to be fairly strong with a linear function being well suited to the model the relationship between the variables.

When assessing associations visually, care must be given. We assess the strength of the association based on how tightly packed the paths are formed by the data. This visual assessment can be greatly distorted by the scale used on the scatter plot. Consider the following scatter plot constructed from the data sets that produced the scatter plots in the first (blue), second (green), and fourth (orange) parts of the previous text exercise.

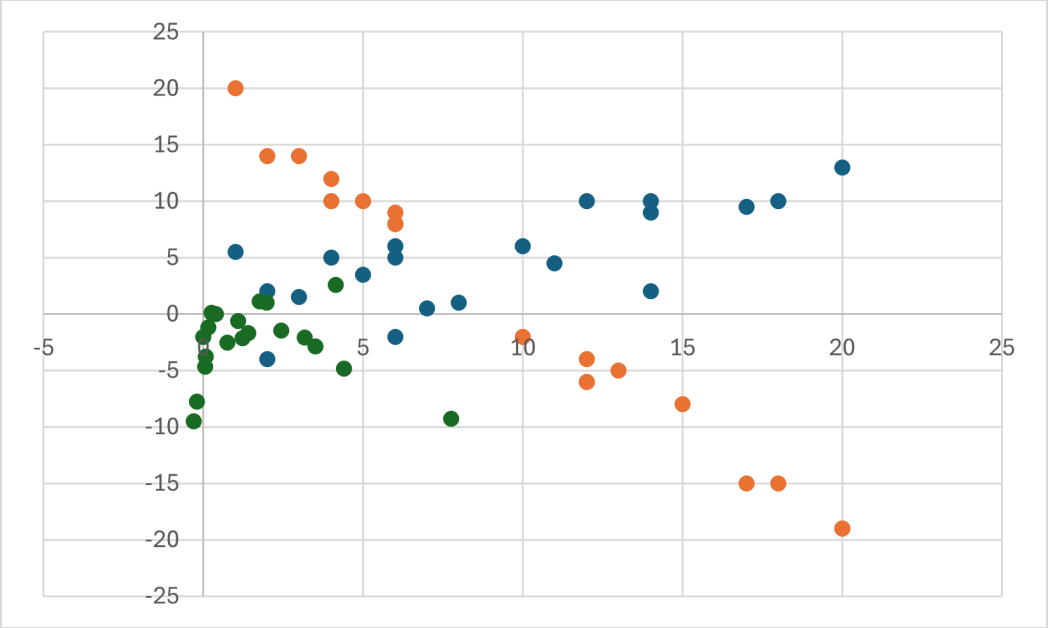


Figure 8.1.12: Comparing scatter plots on same scale

Our understanding of the data from the first and fourth parts of the previous exercise remain the same and in someways are strengthened. The blue data from the first part is positively associated and appears to follow a linear model. The orange data from the fourth part is negatively associated and a linear function still appears to model the data well. Having the two plotted on the same coordinate plane, we can feel confident that the association of the orange data is stronger than that of the blue data because the points form a path that is more densely packed than the blue path. Our confidence in the conclusions relating to the second part of the previous exercise, the green data, may be slightly shaken. At this scale, it seems a little more reasonable to conclude that the green data might be modeled using a linear function with a negative slope. The appearance of a possible linear relationship becomes even more pronounced when the scale is altered again in the following scatter plot.

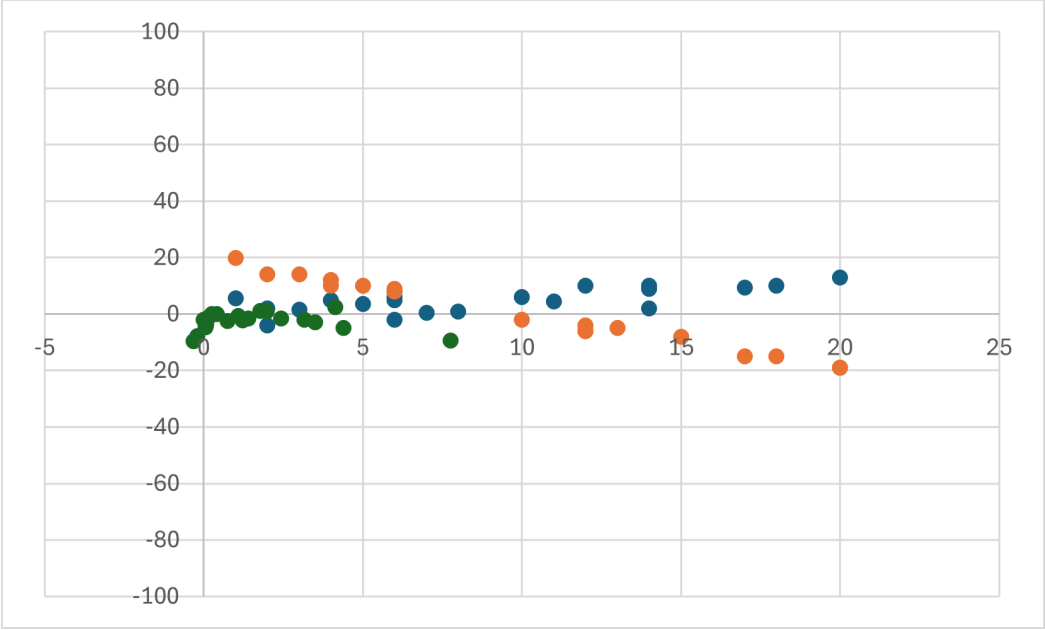


Figure 8.1.13: Comparing scatter plots on same larger scale

Hopefully, we recognize the egregious nature of this last scatter plot. There is no need for such a large scaling of the vertical axis. In the first scatter plot with the three data sets, the scale needed to incorporate all the data present and, therefore, could not be scaled in precisely the same way as the scatter plots were scaled in the original text exercise. Sometimes, there are legitimate

reasons for plotting multiple variables on a single coordinate plane, but when we do so, we must exercise caution. We want to develop analytical methods of assessing the association between variables that does not depend on the scaling of the values either graphically presented or based on the units used in the calculation. We do so later in this chapter.

Correlation and Causation

In this section we address the common adage that "correlation does not imply causation." If quantity A somehow causes or creates quantity B, then one would expect to see an association between the two variables. What the adage is saying, is that the converse is not true; the existence of a correlation could have many explanations other than cause and effect. In reality, the principle applies to more than just correlation; it applies to any association. Perhaps the adage is assembled as an appeal to the allure of alliteration. Setting phrasing aside, the adage warns us that when studying bivariate data, we must carefully discern the conclusions that we draw from the presence of associations.

While the name may be unfamiliar, the concept of causation should be familiar enough; think of it as referencing the ideas of cause and effect. Suppose a pianist begins to play Pachelbel's Canon in D Major at the start of the entrance procession at a wedding. With each press of a key, a hammer strikes a taut string which subsequently reverberates the desired note throughout the space. The pianist, the cause, makes the music, the effect. There is such a possibility when considering quantitative variables as well.

For various reasons, many engaged couples focus on their physiques as part of the preparation for their upcoming wedding. In many cases, this preparation centers around losing weight with a focus on caloric intake. In order to lose weight, one must consistently have a daily caloric deficit. If one wants to gain weight, one must consistently have a daily caloric surplus. Each can be attained through various combinations of food consumption, hydration, and exercise. When in a state of caloric deficiency, the effect is weight loss. We say there is a causal relationship between the quantitative variables caloric deficit and weight loss. When there is a relationship, there is an association. In other words, when there is causation there is association. If we were to collect data on this topic, we would see that the larger the deficit the more weight is lost. We would say that caloric deficit and weight loss are positively correlated. In this case, the positive causal relationship implies a positive correlation.

This is not the typical setup when we are studying the world around us. In general, we do not start with a causal relationship. We start with two variables of interest, collect data, and consider a scatter plot to see if there is an association. Just because there is an association does not mean that the relationship is causal. For example, suppose the happy newly weds are cataloging their gifts so that they may send thank you letters. In doing so, they notice that the people who traveled very far to attend the wedding gave gifts which tended to be more expensive. Their most expensive gifts, in fact, were given by the people who traveled the farthest. Similarly, people who did not travel very far tended to give cheaper gifts. In short, they have noticed an association between two quantitative variables: price of the gift and the distance traveled to the wedding. Is it reasonable to assert that there is a cause and effect relationship? The couple may speculate that when someone puts in a large amount of effort to attend, they are psychologically predisposed to a larger monetary loss and therefore are inclined to spend more than someone who put in little effort to travel. This hypothesis asserts that the travel distance is causing the price of the gift. We wish to emphasize that such a conclusion is premature. While it may be true, there are other explanations of the correlation. Of their loved ones who live far away, perhaps only those with a lot of money were able to attend; those people would be able to afford more expensive gifts. This would suggest that wealth is causing both the willingness to travel long distances and buy expensive gifts. Alternatively, perhaps those who are willing to travel long distances are those who have very close relationships with the couple and thus are willing to spend more money on gifts. Or, it could be a total coincidence; perhaps other weddings did not observe a correlation between these quantities. More work is needed to ascertain which of these explanations, if any, are correct. Concluding any cause and effect based solely on the observation of correlation is erroneous.

In general, if quantity A correlates with quantity B, there are many explanations for that correlation. It could be the case that A causes B. It could be the case that B causes A. It could be the case that a totally separate quantity, C, is causing both A and B; this latter case is referred to as **common response**. Or, as tends to happen with small data sets, the correlation could just be total coincidence: a pattern emerging from random chance and is falsified upon collecting more data. We call these **spurious correlations**. Establishing correlation is comparatively easy, but establishing cause and effect is quite difficult; the latter requires controlled experimentation accompanied with explanatory power. Throughout this chapter, we will explore how to measure correlation, but we will not be making any assertions regarding causation.

8.2: Linear Correlation

Learning Objectives

- Discuss linear correlation
- Motivate and develop Pearson's correlation coefficient r
- Calculate, by hand and using technology, and interpret r
- Emphasize the need for visualization in conjunction with summary statistics

Section 8.2 Excel File (contains all of the data sets for this section)

Review and Preview

When we study bivariate quantitative data, we attempt to identify if there is a relationship between the two variables. If there is a relationship, which we call an association, we would like to describe it as well as we can.

One way to do this is to describe what happens as one of the variables changes. If as one variable increases, the other increases, we say that the variables are positively correlated. If as one variable increases, the other decreases, we say that the variables are negatively correlated. Some associations are not correlated, meaning, as one variable increases, the other variable may increase or decrease depending on which values we are considering. When we have a correlated association, we can further describe the association by determining whether or not the rate of change is constant. In other words, we could determine if a linear function would be a good model for the data. A linear function models the association well when the scatter plot forms a fairly straight path through the coordinate plane. A third way to describe the association is to assess its strength or how densely packed the points are along the path in the scatter plot.

In the previous section, we noted that visual assessments of these considerations can be influenced by the units on the data and the scale of the scatter plots. In this section, we develop an analytical tool to help us, but for best results, we should use both the visualizations and analytical tools in conjunction. As indicated in the previous section, we will restrict our discussion to identifying when an association is well-modeled by a linear function and then measuring the strength of that association. Since linear functions have a constant slope, linear associations are necessarily correlated. As such, we begin our exposition of linear correlation.

Linear Correlation

We are about to embark on the development of Pearson's correlation coefficient, r, for sample data. There is an analogous measure for population data which we will not address in this course. The goal of the correlation coefficient is to assess the strength of the correlation between two variables which are thought to have a linear association. Such a measure would be expected to be independent of both the units describing the data and the ordering of the variables. Returning to our example regarding the ages of the bride and groom on their wedding day, our computation of r should not depend on whether the data is presented in months, years, or decades, and it should not care if our variables are paired as (age of bride, age of groom) or (age of groom, age of bride).

Let us begin with the task of trying to ensure that the measure does not depend on the particular units used to describe the data. Recall that we can use the z-score to compare values within and between different sets of data because the z-score represents how many standard deviations above the mean an observation is. We defined the z-score within the context of population data, but the same concepts apply when we are studying sample data. We are using sample statistics in calculation rather than population parameters. Within this context we have the following.

$$z = \frac{x - \bar{x}}{s}$$

An astute reader will acknowledge a possible issue here. We are now dealing with bivariate data; we have not discussed the idea of means and standard deviations within this context. Luckily, this does not pose any significant issue. We can study each of the individual variables as we have done previously. If we have bivariate data with a first variable x and a second variable y, we can compute the mean and standard deviation of the variable x and then the mean and standard deviation of the variable y in order to compute the associated z-score for each component of each observation. We consider two z-score computations, one for each variable as follows.

$$z_x = \frac{x - \bar{x}}{s_x} \qquad z_y = \frac{y - \bar{y}}{s_y}$$

Text Exercise 8.2.1

1. To begin our exploration, we will consider the following bivariate data set consisting of 10 observations with variables x and y. Construct a scatter plot of bivariate data to confirm that a linear function seems appropriate as a model for the data.

Table 8.2.1: Table of values for variables x and y

x	y
4	100
8	91
12	64
18	46
20	28
22	10
28	10
32	−17
36	−62
40	−80

Answer

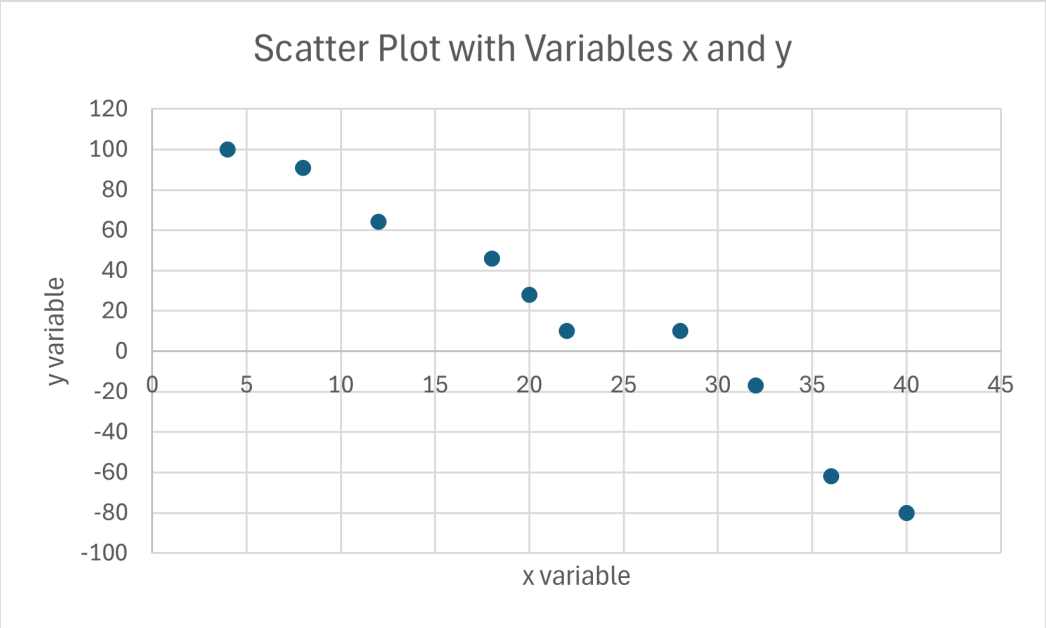


Figure 8.2.1: Scatter plot of variables x and y

After constructing a scatter plot using the variables x and y, we come to the conclusion that it is reasonable to understand the data using a linear function to model the data. It appears that the two variables are negatively correlated.

2. Compute the appropriate z-score for each component of each observation by first computing $\bar{x}$, $\bar{y}$, s_x, and s_y, and then computing the z-score according to which variable the data belongs. Use the following table as a template.

Table 8.2.2: Table of values for variables x and y

x	y	z_x	z_y

x	y	z_x	z_y
4	100		
8	91		
12	64		
18	46		
20	28		
22	10		
28	10		
32	-17		
36	-62		
40	-80		

Answer

In order to compute the appropriate means and standard deviations, we treat each column of our data set as its own separate set of data. We compute $\bar{x}$ by adding all of the values in the x column and dividing by 10 and compute $\bar{y}$ by adding all of the values in the y column and then dividing by 10. We proceed similarly for the computations of the standard deviations by treating the data from each variable separately. Using technology can make this process much quicker. We provide the appropriate values now: $\bar{x} = 22$, $s_x = 12$, $\bar{y} = 19$, and $s_y = 60$. We compute the number of standard deviations each measured value of each variable is away from the mean its particular variable.

Table 8.2.3 Table of values and calculations for variables x and y

x	y	z_x	z_y
4	100	$\frac{4-22}{12} = -\frac{3}{2}$	$\frac{100-19}{60} = \frac{27}{20}$
8	91	$\frac{8-22}{12} = -\frac{7}{6}$	$\frac{91-19}{60} = \frac{6}{5}$
12	64	$\frac{12-22}{12} = -\frac{5}{6}$	$\frac{64-19}{60} = \frac{3}{4}$
18	46	$\frac{18-22}{12} = -\frac{1}{3}$	$\frac{46-19}{60} = \frac{9}{20}$
20	28	$\frac{20-22}{12} = -\frac{1}{6}$	$\frac{18-19}{60} = \frac{3}{20}$
22	10	$\frac{22-22}{12} = 0$	$\frac{10-19}{60} = -\frac{3}{20}$
28	10	$\frac{28-22}{12} = \frac{1}{2}$	$\frac{10-19}{60} = -\frac{3}{20}$
32	-17	$\frac{32-22}{12} = \frac{5}{6}$	$\frac{-17-19}{60} = -\frac{3}{5}$
36	-62	$\frac{36-22}{12} = \frac{7}{6}$	$\frac{-62-19}{60} = -\frac{27}{20}$
40	-80	$\frac{40-22}{12} = \frac{3}{2}$	$\frac{-80-19}{60} = -\frac{33}{20}$

3. Construct a scatter plot of the z-score data and compare it with the scatter plot constructed in the first part of this text exercise.

Answer

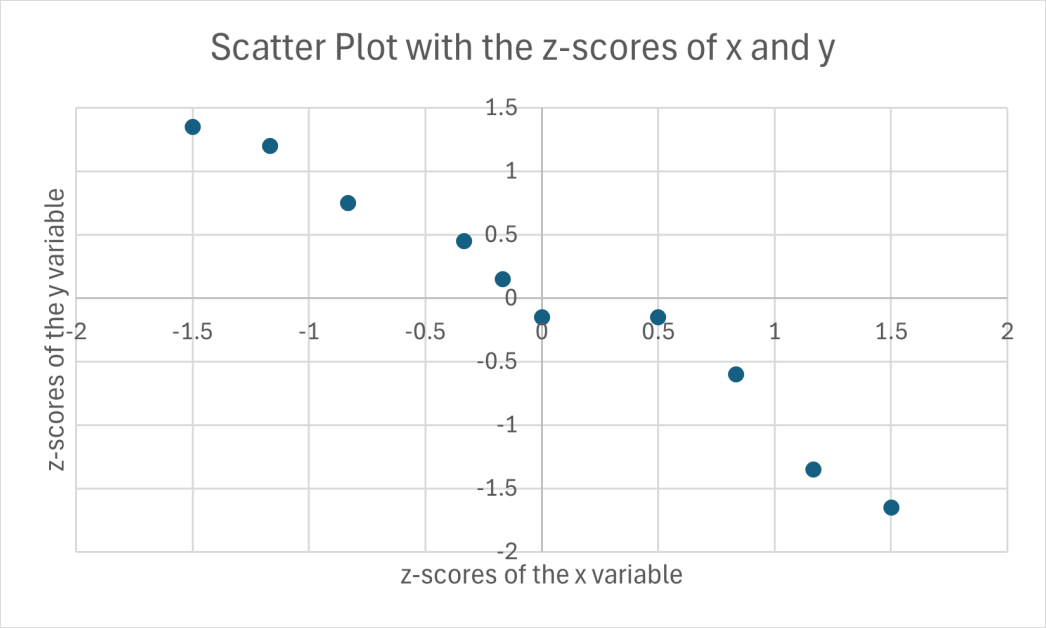


Figure 8.2.2 Scatter plot of z-scores of variables x and y

A visual analysis comparing the scatter plot using the variables z_x and z_y with the scatter plot using the variables x and $y$yields that the shape of the path looks extraordinarily similar. The relative positions of the coordinates seem to match perfectly (which they do). In considering the z-scores of our data, we have successfully removed the impact that different units would make while preserving the relationship in the data.

Let us continue to analyze the scatter plot of the variables z_x and z_y. We note that almost all of the coordinate pairs fall into 2 of the 4 quadrants of the coordinate plane. The top left quadrant, quadrant II, houses 5 of the coordinate pairs; while, the bottom right quadrant, quadrant IV, houses 4 of the coordinate pairs. The last coordinate pair falls on the boundary between quadrants III and IV. We notice that when we have a fairly strong negative correlation the majority of points land in quadrants II and IV.

Text Exercise 8.2.2

1. Recall that the bivariate data that relayed the ages of bride and groom on their wedding day displayed a positive linear correlation when we constructed the scatter plot. Using Excel to transform the data using the z-score, just as we did in the last text exercise, and then construct a scatter plot. Analyze the scatter plot by considering the points in the various quadrants. Does a similar pattern appear?

Answer

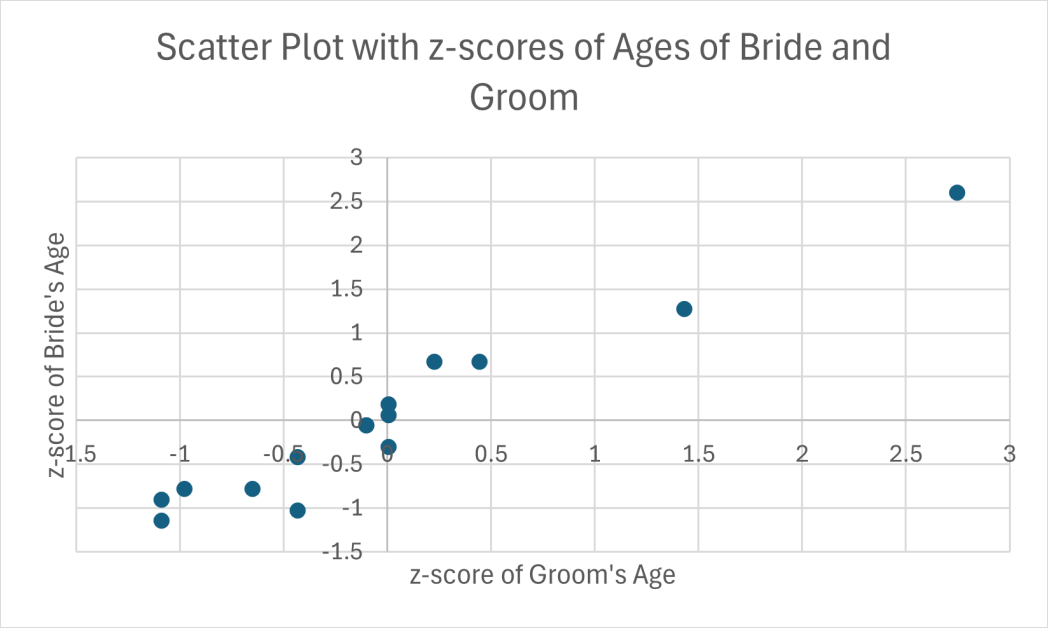


Figure 8.2.3 Scatter plot of z-scores of ages of bride and groom

The scatter plot shows that most of the points fall in quadrant I and quadrant III. There is one point that falls in quadrant IV, but it is very close to the boundary between quadrants III and IV. We see a similar trend to the scatter plot in the previous text exercise that the majority of points fall in two quadrants, but the quadrants are different. Rather than the quadrants labeled with even numbers, we have the quadrants labeled with odd numbers.

2. Now consider the scatter plot of the transformed variables from text exercise 8.1.3.1 presented below. We described the relationship between these variables as having a weaker positive correlation because there was a straight upward path through the data, but the points were not densely packed together. What do you notice about the quadrant analysis in this case?

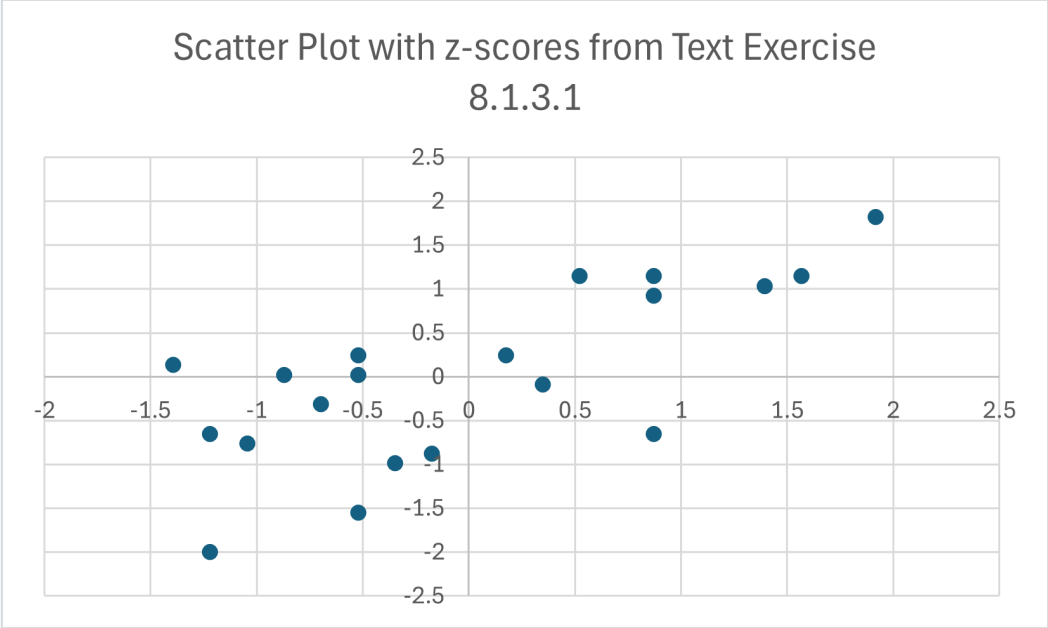


Figure 8.2.4: Scatter plot of z-scores

Answer

The scatter plot reveals points in each of the four quadrants, but the majority of the points fall in the first and third quadrants.

Note: Quadrants with Original Data

We can relate the quadrants back to the original raw data using our understanding of the z-score. Recall that a z-score is positive when the observation is larger than the mean and negative when the observation is smaller than the mean. So, the first quadrant of the scatter plot with z-scores consists of all the points where both variables were greater than the means of their respective variables. The third quadrant consists of all the points where both variables were less than the means of their respective variables, and the second and fourth quadrants consist of the points where one variable was larger and one was smaller. We can build quadrants using the means from our two variables as seen below with the data from text exercise 8.2.1.

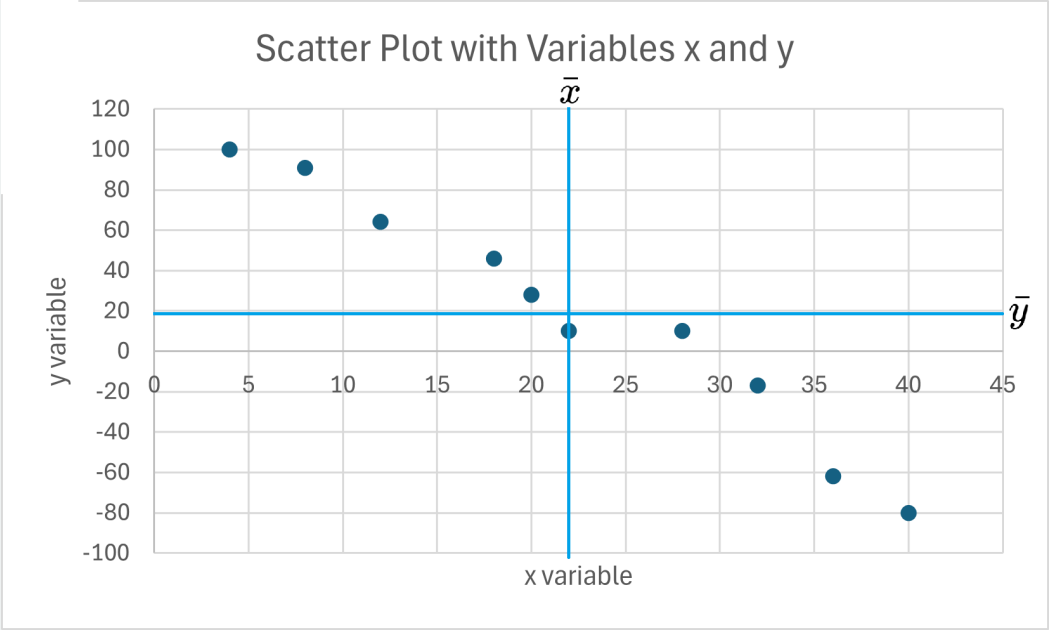


Figure 8.2.5: Scatter plot of variables x and y with $x = \bar{x}$ and $y = \bar{y}$ graphed

We begin to see that the strength of the association is related to the number of points in particular quadrants. With negative associations we see the majority of points in quadrants II and IV; while, we see the majority of points in quadrants I and III when the association is positive. Note that in quadrants I and III, the z-scores have the same sign (either both positive or both negative), and in quadrants II and IV, the z-scores have opposite signs (one negative and one positive). This means that we can readily identify how an observation contributes to the association by the sign of the product its z-scores. If the product is positive, the point falls in either quadrant I or III. If the product is negative, the point falls in either quadrant II or IV. If the majority of the products are positive, we would expect a positive association. If the majority of the products are negative, we would expect a negative association. Of course, there is more at play which we will consider now.

The points on the scatter plot of z-scores that are close to the origin could easily be seen in a path with an upward direction or a downward direction. As such, the number of points that are close to the origin in any given quadrant contribute less to the determination of the association as the points that are farther away from the origin. If the z-scores of an observation have large magnitudes, that observation contributes to the determination of the association to a larger degree. We note that both the sign and the magnitude of the product of an observation's z-scores inform us about the association between the variables. As such, it seems that a reasonable measure of linear correlation would involve averaging all of the products of each observation's z-scores. Indeed, this is what r, the Pearson correlation coefficient, does. We now provide the definition for sample, bivariate, quantitative data coming from a sample size of n with variables x and y.

$$r = \frac{1}{n-1}\sum_{i=1}^{n} z_x \cdot z_y = \frac{1}{n-1}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}$$

Pearson's Correlation Coefficient

Now that we are equipped with the purpose and definition of Pearson's correlation coefficient, we will explore its meaning and implications. Perhaps the first observation made was that the multiplicative factor of $\frac{1}{n-1}$ reminds us of the adjusted averaging that takes place with sample standard deviation. This is for good reason; using such a factor guarantees that the value of r is between -1 and 1 inclusively regardless of the sample size and the magnitudes of the orignal data.

A second observation is that if we switched the variable names in the formula, the result would be exactly the same formula and compute the same value for r; the correlation between x and y is the same as the correlation between y and x. It does not matter which variable gets plotted on the horizontal axis or the vertical axis; the correlation between them will be the same. This was a desired trait mentioned earlier that is achieved with Pearson's correlation coefficient.

As previously discussed, the sign of each product in the computation of r indicates how that particular observation contributes to the association. If r is negative, the summation of the products is negative with the majority of points falling into quadrants II and IV, leading us to conclude that there is a negative correlation. If on the other hand r is positive, the summation of the products is positive which points to the majority of points falling into quadrants I and III leading us to conclude there is a positive correlation. We see that the sign of r tells us the type of correlation present between two variables thought to be linearly associated.

We are not only interested in the type of correlation, but also in the strength of the correlation. If r is 0 or close to 0, some of the products were negative while some of the products were positive; thus, when they were added together, many of the positive values cancelled out many of the negative values resulting in a very small sum. In this case, we expect the points to be distributed fairly evenly across the four quadrants, meaning, the association is weak. If $r = 0$, there is no association. If r is close to 0, there may or may not be some association. Random chance and measurement errors can lead to nonzero r values when there really is no relationship. If there is some association that produced a correlation coefficient close to 0, the weakness of the association, in general, does not warrant further interest. When the r value is 0 or close to 0, we say the variables do not exhibit any correlation.

The only time that r is equal to 1 or -1 is when the association fits a linear function perfectly: every point falls on the same line. If the slope of the line is negative, $r = -1$, and if the slope of the line is positive, $r = 1$. We can assess the strength of a linear association based on how close the value of $|r|$ is to 1.

? Text Exercise 8.2.3

1. Compute, by hand, the the Pearson's correlation coefficient r for the data set examined in text exercise 8.2.1. We replicate the table of values to facilitate the computation. Interpret the meaning of r in light of its sign and magnitude. Compare the findings with what we already know to be true.

Table 8.2.4: Table of values for variables x and y

x	y	z_x	z_y
4	100	$\frac{4-22}{12} = -\frac{3}{2}$	$\frac{100-19}{60} = \frac{27}{20}$
8	91	$\frac{8-22}{12} = -\frac{7}{6}$	$\frac{91-19}{60} = \frac{6}{5}$
12	64	$\frac{12-22}{12} = -\frac{5}{6}$	$\frac{64-19}{60} = \frac{3}{4}$
18	46	$\frac{18-22}{12} = -\frac{1}{3}$	$\frac{46-19}{60} = \frac{9}{20}$
20	28	$\frac{20-22}{12} = -\frac{1}{6}$	$\frac{18-19}{60} = \frac{3}{20}$
22	10	$\frac{22-22}{12} = 0$	$\frac{10-19}{60} = -\frac{3}{20}$
28	10	$\frac{28-22}{12} = \frac{1}{2}$	$\frac{10-19}{60} = -\frac{3}{20}$
32	-17	$\frac{32-22}{12} = \frac{5}{6}$	$\frac{-17-19}{60} = -\frac{3}{5}$
36	-62	$\frac{36-22}{12} = \frac{7}{6}$	$\frac{-62-19}{60} = -\frac{27}{20}$
40	-80	$\frac{40-22}{12} = \frac{3}{2}$	$\frac{-80-19}{60} = -\frac{33}{20}$

Answer

Table 8.2.5 Table of values and computations for variables x and y

x	y	z_x	z_y	$z_x \cdot z_y$
4	100	$\frac{4-22}{12} = -\frac{3}{2}$	$\frac{100-19}{60} = \frac{27}{20}$	$-\frac{3}{2} \cdot \frac{27}{20} = -\frac{81}{40}$
8	91	$\frac{8-22}{12} = -\frac{7}{6}$	$\frac{91-19}{60} = \frac{6}{5}$	$-\frac{7}{6} \cdot \frac{6}{5} = -\frac{7}{5}$
12	64	$\frac{12-22}{12} = -\frac{5}{6}$	$\frac{64-19}{60} = \frac{3}{4}$	$-\frac{5}{6} \cdot \frac{3}{4} = -\frac{5}{8}$
18	46	$\frac{18-22}{12} = -\frac{1}{3}$	$\frac{46-19}{60} = \frac{9}{20}$	$-\frac{1}{3} \cdot \frac{9}{20} = -\frac{3}{20}$
20	28	$\frac{20-22}{12} = -\frac{1}{6}$	$\frac{18-19}{60} = \frac{3}{20}$	$-\frac{1}{6} \cdot \frac{3}{20} = -\frac{1}{40}$
22	10	$\frac{22-22}{12} = 0$	$\frac{10-19}{60} = -\frac{3}{20}$	$0 \cdot -\frac{3}{20} = 0$
28	10	$\frac{28-22}{12} = \frac{1}{2}$	$\frac{10-19}{60} = -\frac{3}{20}$	$\frac{1}{2} \cdot -\frac{3}{20} = -\frac{3}{40}$
32	-17	$\frac{32-22}{12} = \frac{5}{6}$	$\frac{-17-19}{60} = -\frac{3}{5}$	$\frac{5}{6} \cdot -\frac{3}{5} = -\frac{1}{2}$
36	-62	$\frac{36-22}{12} = \frac{7}{6}$	$\frac{-62-19}{60} = -\frac{27}{20}$	$\frac{7}{6} \cdot -\frac{27}{20} = -\frac{63}{40}$
40	-80	$\frac{40-22}{12} = \frac{3}{2}$	$\frac{-80-19}{60} = -\frac{33}{20}$	$\frac{3}{2} \cdot -\frac{33}{20} = -\frac{99}{40}$

Having computed all of the products of the z-scores in the far right column, all we need to do is sum them up and divide by 9.

$$r = \frac{1}{10-1} \sum_{i=1}^{10} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y} = \frac{1}{9} \cdot -\frac{177}{20} = -\frac{59}{60} = -0.983\bar{3}$$

The negative value of r indicates a negative correlation between the variables x and y. The magnitude of 0.9833which is close to 1 indicates that the linear association between x and y is quite strong. The understanding derived from our considerations of the correlation coefficient r align with the conclusions previously drawn.

2. Consider the following bivariate data set using both a scatter plot and r.

Table 8.2.6: Table of values for variables x and y

x	y
-2	4
0	4
1	4
2	4
5	4

Answer

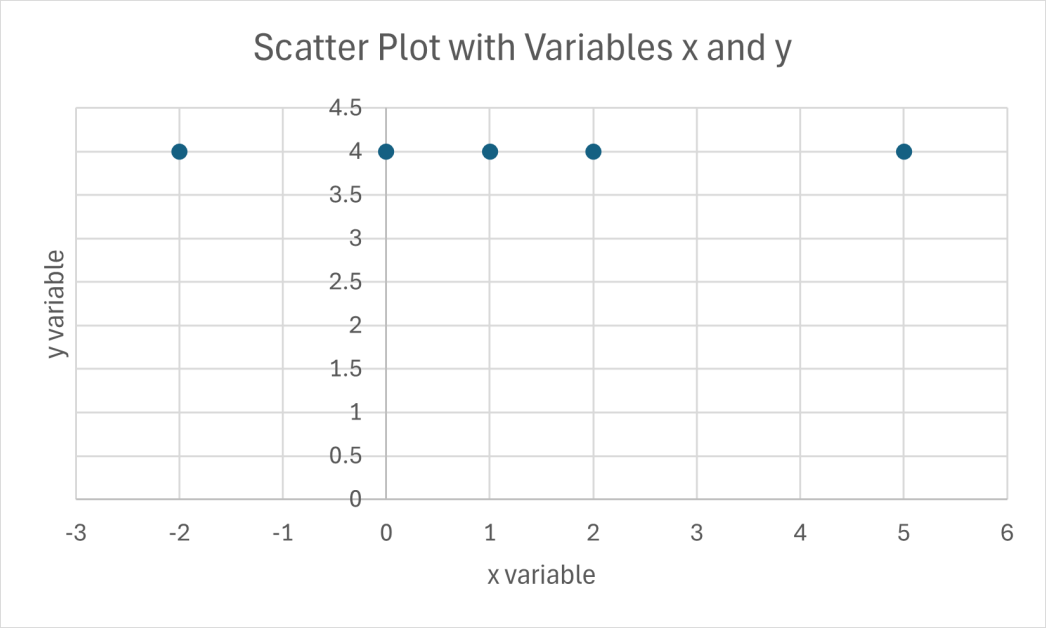


Figure 8.2.6 Scatter plot

The scatter plot shows points that form a perfectly straight horizontal line. When computing the correlation coefficient, we run into problems, because the standard deviation of the variable y is 0. We cannot divide by 0. Since every y value is 4, the mean is 4, and the standard deviation is 0. As such, it appears that the correlation coefficient fails to recognize a perfectly linear association. In reality this is not the case. When our data aligns in either a perfect horizontal line or vertical line, there is no association between the variables. There is no relationship between the value of the x variable and the value of the y variable because one of the variables is fixed regardless of the other variable. Since there is no relationship between the variables, it does not make sense to measure the degree of a linear relationship.

As we saw with a data set containing just 10 observations, the computation of the correlation coefficient can be quite tedious. As such, the computation is often carried out using technology. The function within Excel that computes Pearson's correlation coefficient r is the CORREL function. The function takes, as inputs, two arrays of numbers that are the same size. The first array consists of all the values of one of the variables; the second array consists of all the values of the other variable. The program matches the arrays by position to pair the values of each variable.

3. Using technology, reconstruct the scatter plot of the previously considered data and calculate Pearson's correlation coefficient r.

Table 8.2.7: Age of bride and room on wedding day

Groom's Age (years)	Bride's Age (years)
20	21
26	20
32	34
30	30
21	22
29	28
26	25
34	34
29	28

Groom's Age (years)	Bride's Age (years)
55	50
30	26
43	39
30	29
24	22
20	19

Answer

$r = 0.9689$

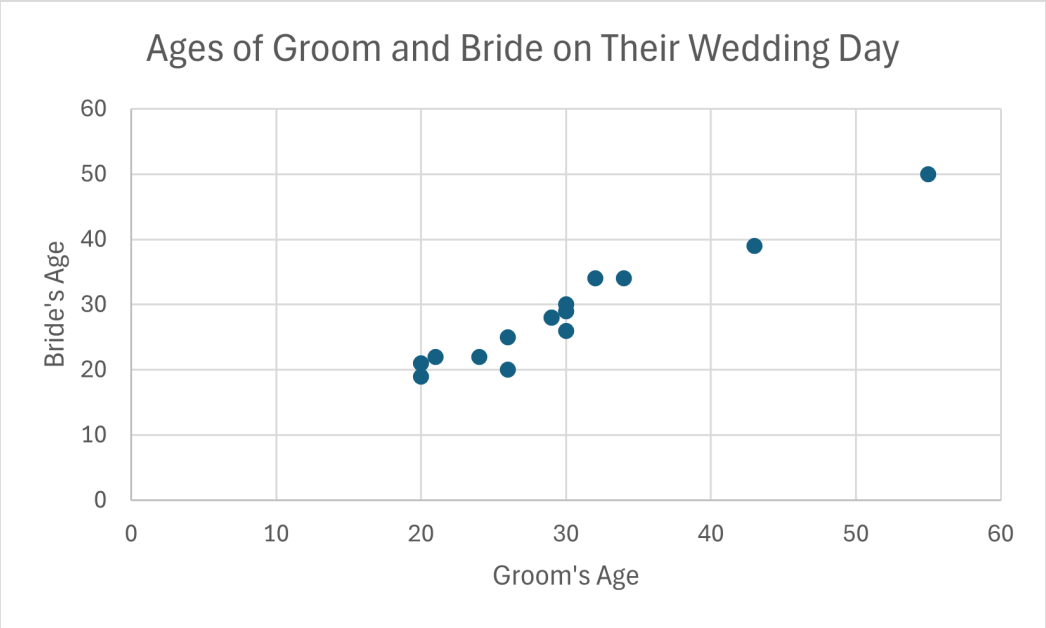


Figure 8.2.7: Scatter plot of ages of bride and groom

4. Using technology, compute Pearson's correlation coefficient r.

Table 8.2.8: Age of bride vs number of children

Bride's Age (years)	Number of Children
19	2
21	8
21	5
22	0
22	3
23	6
23	4
23	2
24	7

Bride's Age (years)	Number of Children
25	1
25	3
26	3
27	4
29	5
31	3
31	1
32	2
33	0
35	2
41	1

Answer

$r = -0.4247$

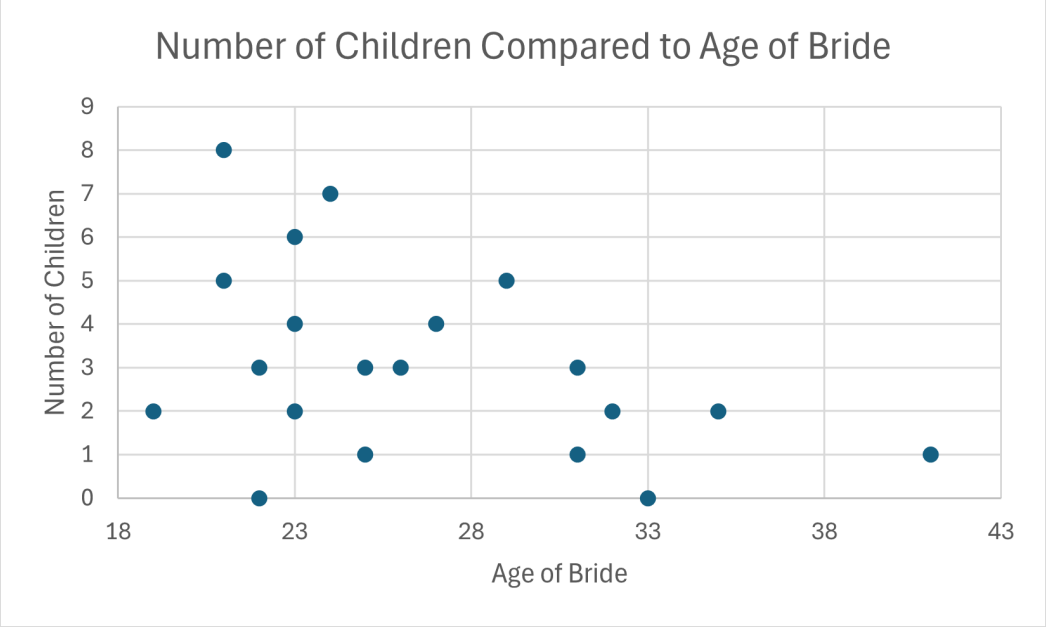


Figure 8.2.8 Scatter plot of age of bride and number of children

5. Using technology, compute Pearson's correlation coefficient r.

Table 8.2.9: Bivariate data from Text Exercise 8.1.3.1

x	y
4	5
11	4.5
10	6
14	10

x	y
6	6
6	-2
2	2
8	1
12	10
7	0.5
14	9
5	3.5
3	1.5
17	9.5
1	5.5
18	10
14	2
20	13
2	-4
6	5

Answer

$r = 0.7253$

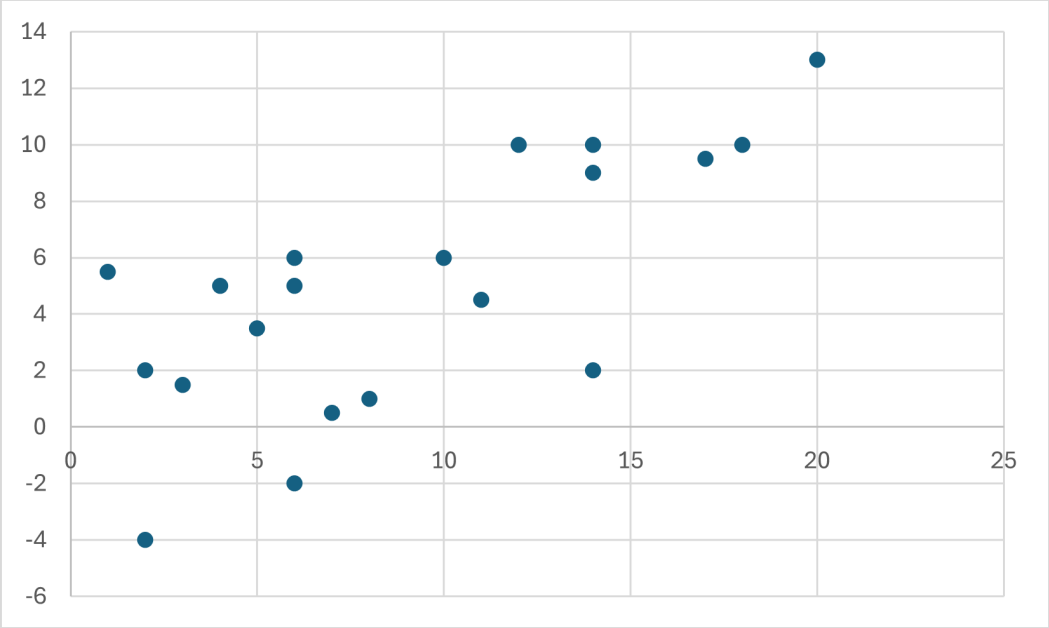


Figure 8.2.9 Scatter plot

6. Using technology, compute Pearson's correlation coefficient r.

Table 8.2.10: Bivariate data from Text Exercise 8.1.3.4

x	y
12	-6
10	-2
12	-6
6	9
20	-19
15	-8
17	-15
3	14
6	8
18	-15
5	10
1	20
12	-4
6	8
2	14
20	-19
6	8
4	10
4	12
13	-5

Answer

$r = -0.9941$

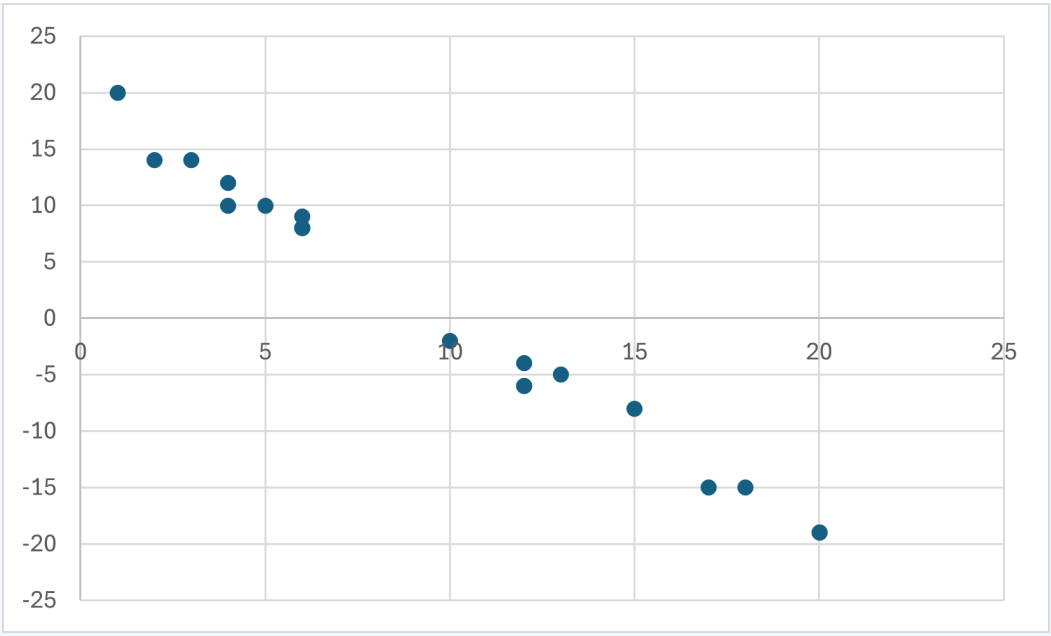


Figure 8.2.10 Scatter plot

Text Exercise 8.2.4

For each of the following scatter plots, determine if the proposed r value is reasonable. If not, explain why.

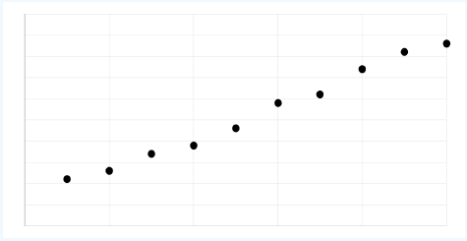
1. 

Figure 8.2.11: Scatter plot

$r = 0.99$

Answer

The scatter plot displays strong positive correlation that appears very close to a line. As such, a positive value close to 1 seems reasonable.

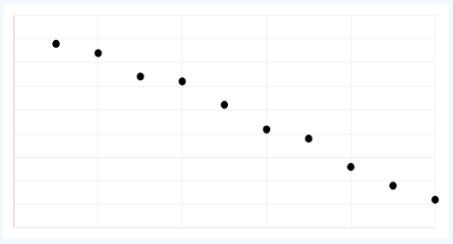
2. 

Figure 8.2.12: Scatter plot

$r = 0.95$

Answer

The scatter plot displays strong negative correlation that appears very close to a line. As such, the proposed r value is unreasonable. The magnitude might be reasonable, but it is clear that the r value should be negative.

3. 

Figure 8.2.13: Scatter plot

$$r = 0.85$$

Answer

The scatter plot does not display much correlation. As such, a value of 0.85 seems unreasonably high.

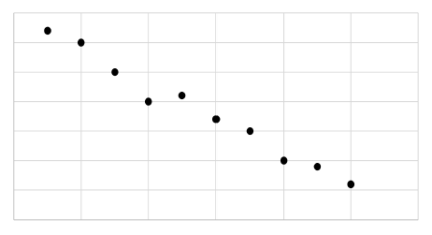
4. 

Figure 8.2.14: Scatter plot

$$r = -1$$

Answer

The scatter plot displays a strong negative linear correlation. However, we can easily see that the data does not fit perfectly on a line. As such, the proposed r value is unreasonable.

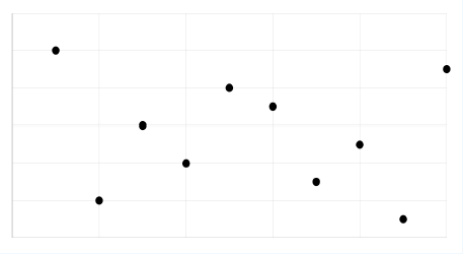
5. 

Figure 8.2.15: Scatter plot

$$r = -0.15$$

Answer

The scatter plot does not display much correlation. As such, an r value close to 0 seems reasonable. Given the general downward direction of the data, a negative r value seems reasonable. We thus conclude that such an r value seems reasonable.

Text Exercise 8.2.5

1. The statistician Francis Anscombe constructed 4 data sets in 1973 that have earned the moniker Anscombe's Quartet. In this text exercise, we will analyze each of the data sets individually and then consider them together. For each of the data sets, compute $\bar{x}$, s_x, $\bar{y}$, s_y, and r. What similarities are there between the data sets? What conclusions can be drawn?

Table 8.2.11: Anscombe's Quartet

Data Set I (x,y)		Data Set II (x,y)		Data Set III (x,y)		Data Set IV (x,y)	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Answer

Data Set I: $\bar{x} = 9$, $s_x \approx 3.3166$,$\bar{y} \approx 7.5009$,$s_y \approx 2.0316$,and $r \approx 0.8164$

Data Set II: $\bar{x} = 9$, $s_x \approx 3.3166$,$\bar{y} \approx 7.5009$,$s_y \approx 2.0317$,and $r \approx 0.8162$

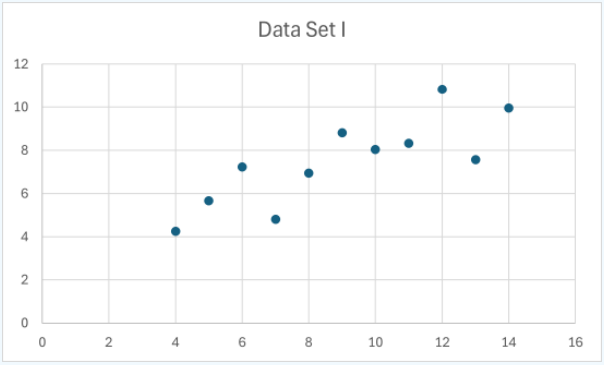
Data Set III: $\bar{x} = 9$, $s_x \approx 3.3166$,$\bar{y} \approx 7.5$, $s_y \approx 2.0304$,and $r \approx 0.8163$

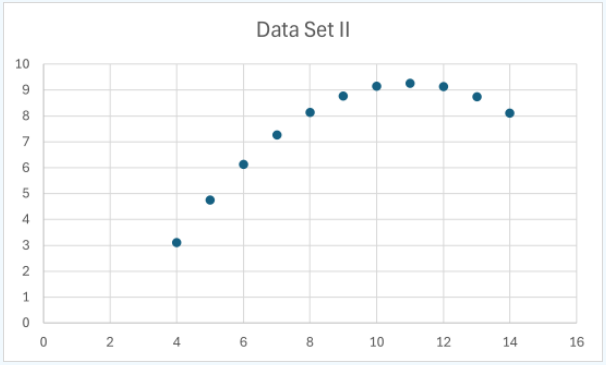
Data Set IV: $\bar{x} = 9$, $s_x \approx 3.3166$,$\bar{y} \approx 7.5009$,$s_y \approx 2.0306$,and $r \approx 0.8165$

The summary statistics for each of the data sets are remarkably similar. They all match up to two or three decimal places. From the perspective of these summary statistics, the data sets are almost indistinguishable.

2. Having computed the summary statistics for each data set, construct scatter plots for each of the data sets. What conclusions can now be drawn? What implications does this exercise have on conducting statistical analyses?

Answer





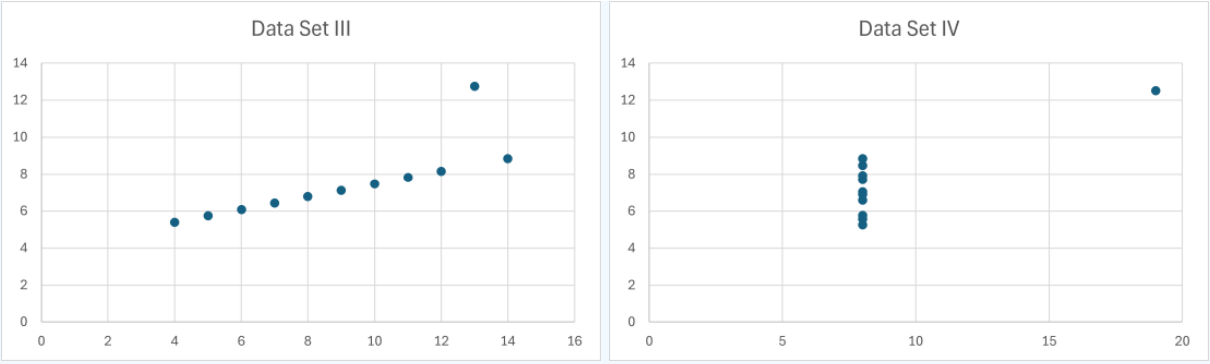


Figure 8.2.16 Scatter plots of Anscombe's Quartet

The scatter plots provide crucial information regarding the various data sets. The first scatter plot reveals an apparent linear relationship with a fairly decent association (the r value confirms this). The second scatter plot reveals a nonlinear relationship indicating that using the correlation coefficient to describe the relationship should not have been done. Indeed, for this particular nonlinear relationship, it appears for some values the association is positive, but for others the association is negative. Thus measuring for any correlation, linear or not, is not to be done. The third scatter plot reveals a quite strong linear relationship with an apparent outlier which resulted in the correlation coefficient dropping significantly in value. The fourth scatter plot reveals the presence of an outlier. The remainder of the data seems to indicate that there is no relationship between the variables.

The visual representations of data are crucial components of statistical analysis. The scatter plots provide us with the information to determine if it is even reasonable to use the correlation coefficient as a measure. When conducting statistical analyses, constructing visualizations is generally the first step, as they provide much information and a general feel for the data. A final takeaway is that we do not want to blindly compute the correlation coefficient and use it as a measure for the presence of a linear relationship. It is best used as a way to measure the strength of a linear relationship that is thought to exist based on scatter plots or other lines of reasoning.

8.3: Introduction to Simple Linear Regression

Learning Objectives

- Define independent and dependent variables
- Differentiate between observed and predicted values of the dependent variable
- Compare different linear models using the sum of squared errors
- Motivate the line of best fit and provide means for its computation
- Develop the coefficient of determination
- Conduct linear regression analysis by checking the reasonability of using a linear model, computing R^2, and finding the line of best fit
- Nuance the predictive and interpretive power of linear regression

Section 8.3 Excel File: (contains all of the data sets for this section)

Review and Preview

In studying bivariate quantitative data, we try to determine whether there is an association between two particular variables or not. If there is an association, a relationship between the variables, we would like to describe the relationship. We are interested in what happens to one variable as the other variable changes. We have discussed several ways to build this understanding: constructing scatter plots, classifying associations, and determining correlation. While more advanced textbooks address nonlinear correlation, we restricted ourselves to linear correlation. Linear correlation assesses the strength of an underlying linear relationship between the two variables of interest. If there is a linear relationship, it seems appropriate to think that there is a linear function that models the relationship. Knowledge of such a function would deepen our understanding of the relationship and allow us to extrapolate regarding values that were not explicitly measured in our collection of the data. In essence, such a function would enable us to make predictions about cases that were not explicitly studied. One of the fundamental motivations of statistical inquiry is to understand the world better so that we may better predict what will happen and act accordingly. This section develops the ideas of constructing a linear function that is the best fit for the data at hand.

When there is a linear relationship between two variables x and y, we expect there to be constants m and b such that $y = mx + b$. In this formulation, we refer to y (the vertical variable) as the **dependent variable** and x (the horizontal variable) as the **independent variable**. Recall that m is the slope of the line, the amount of change in y if x is increased by 1, and b is the y-intercept, the y value of the line when the x value is 0. Since there is a relationship between the two variables, one variable changes as the other changes; that is, the slope of the line is defined and not 0; we have $m \neq 0$.

Remember that our study of correlation did not depend on the ordering of the variables. If there is a linear relationship between x and y, there is a linear relationship between y and x. In which case, we would expect another set of constants say M and B such that $x = My + B$. We would call x the dependent variable and y the independent variable. When studying associations, we do not assume causal relationships; do not let the terminology influence your thought in this regard.

When we collect data to understand the relationship, we expect the data to have some natural variation from the equation due to measurement error, natural variation, and the random noise that occurs in reality. As such, when we use collected data to construct a linear function, we are approximating the values of m and b. Throughout this section, we will use the hat symbol to indicate approximation values. Thus m is approximated by $\hat{m}$ and b is approximated by $\hat{b}$. We will use these values to approximate y values which we denote $\hat{y}$ using the following equation.

$$\hat{y} = \hat{m}x + \hat{b}$$

Several fundamental questions arise. How do we pick the values of $\hat{m}$ and $\hat{b}$? How do we know how well we did in picking the values of $\hat{m}$ and $\hat{b}$? Is there an optimal choice of values for $\hat{m}$ and $\hat{b}$? Even if we have the best line, it will not be perfect unless all the points are on the same line. Given that there is some error, how well does the line fit the data? These are questions we will begin to answer. We call the process of finding and evaluating these lines **regression analysis**.

Modeling Using Linear Functions

When studying bivariate quantitative data, we do not expect, even when there is a linear relationship, that all of the data points fall precisely on the same line. As such, even once we decide upon a linear function to model the data, it is impossible for the function to match the data perfectly. There will necessarily be some error between some of the observed values and the predicted values associated with them. To visualize this consider the following data set along with a scatter plot visualizing it.

Table 8.3.1: Paired values of variables x and y

Observation	x	y
1	7	5
2	2	4
3	6	6
4	3	3
5	1	2

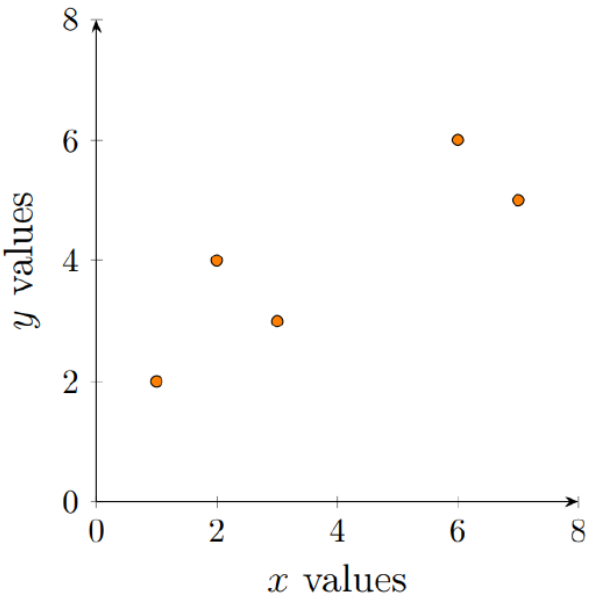


Figure 8.3.1: Scatter plot of x and y

Notice that we cannot draw a line that goes through every point of the scatter plot. We could fairly easily plot a line through 2 of the points or even 3 of the points, but we cannot go through all 5 points; we cannot avoid the presence of error in our model! Suppose that we decided upon the linear function $\hat{y} = 0.5x + 2$ to model this particular data set and then plotted it below in dark blue. Notice that the line does not go through any of the observed values plotted on the scatter plot. The observed values remain in orange, and the predicted values are colored in a light blue.

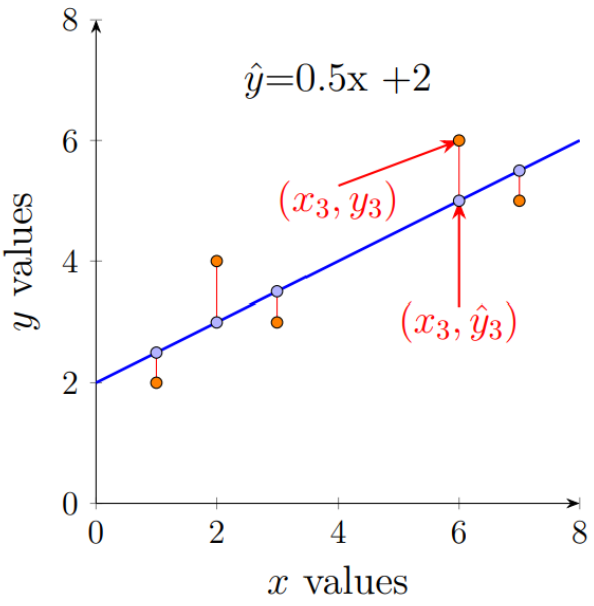


Figure 8.3.2: Scatter plot with linear model

The scatter plot labels two particular coordinate pairs on the scatter plot: $(x_3, y_3) = (6, 6)$ and $(x_3, \hat{y}_3) = (6, 5)$. The former is the observed pair; while, the latter is the predicted pair. Notice how the x-coordinates are the same. The predicted value of y_3 is $\hat{y}_3$. We compute $\hat{y}_3$ using the equation of the line $\hat{y}_3 = 0.5x_3 + 2 = 0.5 * 6 + 2 = 5$. We call the difference between y_3 and $\hat{y}_3$ the error at x_3 which we denote e_3. In general, we define the error at any observed x value x_i as follows.

$$e_i = y_i - \hat{y}_i$$

? Text Exercise 8.3.1

1. Using the linear function $\hat{y} = 0.5x + 2$, predict the value of y when $x = 8$.

Answer

We can predict the value of the variable y be evaluating the linear function at the indicated x value. We have a predicted value of $0.5 \cdot 8 + 2 = 6$.

2. Using the linear function $\hat{y} = 0.5x + 2$, compute the error at each of the 5 collected x values.

Table 8.3.2: Values for variables x and y

Observation	x	y	$\hat{y}$	$e_i = y - \hat{y}$
1	7	5		
2	2	4		
3	6	6		
4	3	3		
5	1	2		

Answer

The error is computed by taking the difference between the y value and the $\hat{y}$ value at a given x value. We must determine each $\hat{y}$ value. We do so in the table that follows.

Table 8.3.3 Computation of predicted values and errors

Observation	x	y	$\hat{y}$	$e_i = y - \hat{y}$
1	7	5	$0.5 \cdot 7 + 2 = 5.5$	$5 - 5.5 = -0.5$
2	2	4	$0.5 \cdot 2 + 2 = 3$	$4 - 3 = 1$
3	6	6	$0.5 \cdot 6 + 2 = 5$	$6 - 5 = 1$
4	3	3	$0.5 \cdot 3 + 2 = 3.5$	$3 - 3.5 = -0.5$
5	1	2	$0.5 \cdot 1 + 2 = 2.5$	$2 - 2.5 = -0.5$

Notice that when the observed value is above the linear function modeling the data that the error is positive and when the observed value is below the line the error is negative. With this basis, we begin to develop the process of determining the best fitting line.

The Line of Best Fit

When we gather information about the world around us, we collect a lot of information. In order to understand best, we try to incorporate as much of the data as we can in our analyses and considerations. We do not collect samples from many people and then only use the results from a handful. Each observation provides important information; we do not want to exclude information without due cause. So, how do we decide upon a line to model our data, when no line perfectly predicts our data in practice? We want to use all of the data in the construction of the line, but how do we achieve such a goal when a line is uniquely determined given two points or a point and a slope?

The answer resides in considering error. We could assess the quality of a line by looking at the error across all observed values, but how are we to assess the totality of the error? If we were sum or average the errors, we would get cancelation between positive and negative errors. Indeed, if we modeled the data from above with the constant function $\hat{y} = 4$, the sum of the errors and hence the average would be 0, but the function would only go through one point and would indicate that there is no relationship between the variables. This does not match with the reality of the data. It is natural to desire that the measure of the totality of error is 0 only when the model perfectly fits the data. We, therefore, must expand our considerations.

Hopefully, we remember a similar discussion surrounding how to measure the dispersion of a data set. We went through several possibilities until we settled on our definition of variance which involved summing the deviations from the mean squared. We will utilize a similar sort of methodology without much motivation; we will consider the **sum of the squared errors** (**SSE**) as our measure of the totality of the error present in the model. In setting this as our measure, we have that there exists a unique line that minimizes the SSE. We can thus find a line and assert that it is one and only best line. We refer to this line as the line of best fit or the least-squares regression line.

? Text Exercise 8.3.2

1. Compute the SSE using the function $\hat{y} = 0.5x + 2$ on the same data set as before, which we reproduce below.

Table 8.3.4: Values for variables x and y

Observation	x	y	$\hat{y}$	$e_i = y - \hat{y}$	$e_i^2 = (y - \hat{y})^2$
1	7	5	5.5	$5 - 5.5 = -0.5$	
2	2	4	3	$4 - 3 = 1$	
3	6	6	5	$6 - 5 = 1$	
4	3	3	3.5	$3 - 3.5 = -0.5$	
5	1	2	2.5	$2 - 2.5 = -0.5$	

Answer

We have computed the error for each of the observed x values in a previous text exercise. All that is left to do is square each of the errors and then add them together.

Table 8.3.5 Computation of predicted values, errors, and squared errors

Observation	x	y	$\hat{y}$	$e_i = y - \hat{y}$	$e_i^2 = (y - \hat{y})^2$
1	7	5	5.5	$5 - 5.5 = -0.5$	$(-0.5)^2 = 0.25$
2	2	4	3	$4 - 3 = 1$	$1^2 = 1$
3	6	6	5	$6 - 5 = 1$	$1^2 = 1$
4	3	3	3.5	$3 - 3.5 = -0.5$	$(-0.5)^2 = 0.25$
5	1	2	2.5	$2 - 2.5 = -0.5$	$(-0.5)^2 = 0.25$

The sum of the squared errors is thus SSE $= 0.25 + 1 + 1 + 0.25 + 0.25 = 2.75$.

2. We can show that the function $\hat{y} = 0.5x + 2$ is not the best fitting line by finding another line that has a smaller SSE. One of the properties of the line of best fit is that it goes through the point $(\bar{x}, \bar{y})$. Let us keep the same slope but adjust the y-intercept so that our function goes through the point $(\bar{x}, \bar{y})$ and then compute the SSE.

Table 8.3.6: Values for variables x and y

Observation	x	y	$\hat{y}$	$e_i^2 = (y - \hat{y})^2$
1	7	5		
2	2	4		
3	6	6		
4	3	3		
5	1	2		

Answer

We are going to model our data with the function $\hat{y} = 0.5x + b$ so that our function passes through the point $(\bar{x}, \bar{y})$. In order to compute b, we must compute $\bar{x}$ and $\bar{y}$, which we find to be $\bar{x} = 3.8$ and $\bar{y} = 4$. Knowing the slope and a point through which our line passes is enough to determine the value of b. We plug the x- and y-coordinates into the function and solve for b.

$$\begin{aligned} 4 &= 0.5 \cdot 3.8 + b \\ 4 &= 1.9 + b \\ 4 - 1.9 &= b \\ 2.1 &= b \end{aligned}$$

Thus producing the function $\hat{y} = 0.5x + 2.1$ as our linear model for the data set. From this we can produce a table of the approximated values at each observed x value and then compute each of the squared errors.

Table 8.3.7 Computation of predicted values and squared errors

Observation	x	y	$\hat{y}$	$e_i^2 = (y - \hat{y})^2$
1	7	5	5.6	$(-0.6)^2 = 0.36$
2	2	4	3.1	$0.9^2 = 0.81$
3	6	6	5.1	$0.9^2 = 0.81$
4	3	3	3.6	$(-0.6)^2 = 0.36$
5	1	2	2.6	$(-0.6)^2 = 0.36$

The sum of the squared errors is thus SSE $= 0.36 + 0.81 + 0.81 + 0.36 + 0.36 = 2.7$. Notice how the individual errors increased in magnitude for some values but decreased for others when comparing the error values using the previous model with these error values. Also, note that the sum of the errors is 0, again emphasizing the inadequacy of using the sum or average of the errors as a measure of the totality of error. We finally notice that the SSE is smaller with this new model than with the old. We thus say that this model is better than the previous model. The question is, have we found the best model yet?

3. Using technology, we determined that the line of best fit for the data set in question is given by the equation $\hat{y} = .5224x + 2.0149$ where the values of $\hat{m}$ and $\hat{b}$ are rounded to four decimal places. Compute the SSE.

Table 8.3.8: Values for variables x and y

Observation	x	y	$\hat{y}$	$e_i^2 = (y - \hat{y})^2$
1	7	5		
2	2	4		
3	6	6		
4	3	3		
5	1	2		

Answer

We have computed the error for each of the observed x values in a previous text exercise. All that is left to do is square each of the errors and then add them together.

Table 8.3.9 Computation of predicted values and squared errors

Observation	x	y	$\hat{y}$	$e_i^2 = (y - \hat{y})^2$
1	7	5	5.6717	$(-0.6717)^2 \approx 0.4512$
2	2	4	3.0597	$0.9403^2 \approx 0.8842$
3	6	6	5.1493	$0.8507^2 \approx 0.7237$
4	3	3	3.5821	$(-0.5821)^2 \approx 0.3388$
5	1	2	2.5373	$(-0.5373)^2 \approx 0.2887$

The sum of the squared errors is thus SSE ≈ 2.6866. We again note that this is the smallest of the sum of squared errors that we have yet to see for this data set. Indeed, this is the smallest attainable value for any linear function! No matter what slope and y-intercept picked, the sum of the squared errors will be larger than this value. Readers with a background in calculus or linear algebra are encouraged to read or work out the details for why this is!

Establishing the uniqueness and computational formulas for the line of best fit requires mathematics beyond the scope of this course. The mathematics does, however, produce very elegant results regarding the slope and y-intercept for the line of best fit, the line that minimizes the sum of the squared errors. We provide the formulas without proof.

$$\hat{m} = r\frac{s_y}{s_x} \qquad \hat{b} = \bar{y} - r\frac{s_y}{s_x}\bar{x} = \bar{y} - \hat{m}\bar{x}$$

? Text Exercise 8.3.3

Using the coefficients for the line of best fit provided above, show that every line of best fit, regardless of the data set, goes through the point $(\bar{x}, \bar{y})$. That is, show that when you substitute $\bar{x}$ in the formula for the line of best fit, the value returned is $\bar{y}$.

Answer

The line of best fit is given by the formula $\hat{y} = \hat{m}x + \hat{b}$. We were given the formulas for $\hat{m}$ and $\hat{b}$. We have $\hat{y} = r\frac{s_y}{s_x}x + \bar{y} - r\frac{s_y}{s_x}\bar{x}$. When we substitute $\bar{x}$ into the variable x, we obtain the following.

$$\begin{aligned}\hat{y} &= r\frac{s_y}{s_x}\bar{x} + \bar{y} - r\frac{s_y}{s_x}\bar{x} \\ &= \bar{y}\end{aligned}$$

In addition to the fact that the point $(\bar{x}, \bar{y})$ always falls on the line of best fit, we have the sum and the average of all the errors is always 0 (when we do not round the numbers). We will not provide a proof of such a fact here, but it is a fact worth noting.

Just as with the computation of the correlation coefficient, we generally rely on technology to compute the slope and y-intercept for the line of best fit. We provide you with the function in Excel that returns the desired information as an array with the slope in the left cell and the y-intercept in the right cell. The function name is LINEST and takes four arguments: the first is the array of y values (values of the dependent variable); the second is the array of x values (values of the independent variable); the third is set of TRUE or 1; and the fourth is set to FALSE or simply 0. Further information is available using the fourth argument but will not be utilized in this course.

Assessing the Line of Best Fit

We have now established that we can find the line of best fit, but another consideration must be made. Just because something is the best does not necessarily mean it is good. Of all the lines that could be used to model the data, we can find the best one, but does this best line actually fit the data well? This is the question we seek to answer and seems closely related to the correlation coefficient. Since the correlation coefficient measures the strength of an apparent linear relationship, we would expect that the closer $|r|$ is to 1, the better our line of best fit will model the data. This intuition is correct and will be confirmed as we approach the problem from a different direction.

When we are studying bivariate quantitative data (variables x and y) we are interested in how one variable changes as the other changes. With this, we may ask how much of the change in one variable can be attributed to the change of the other variable? Inherently, this question requires the development of some method or model which can measure the amount of change in the dependent variable which can be attributed to the model. When making such a measurement, the interest lies in the proportion of the change in one variable that can be attributed to the model, not the raw amount of variation that can be attributed. This allows the measure to be compared across data sets composed of data with vastly different magnitudes and makes the measure value independent of the units of the measurement. A high percentage indicates that the model fits well. Most of the change in y can be explained as due to the change in the x variable. If the percentage is low, the model does not fit well. The majority of the change in y is not understood as due to changes in x under the model.

In order to continue, we must decide how to measure the change in the y variable; this is really a question of dispersion. In general, the more the y variable changes, the greater the spread of the y variable data. Our most commonly used measure of dispersion has been standard deviation, but as we have seen throughout our bonus discussions, the real statistical power lies not in standard deviation but in variance. Recall that the variance is closely related to the average of square deviations from the mean, but we are not interested in a typical value, rather, we want the total change in the y variable. As such, we define our measure of change in y to be the **total variation** of y which we can compute with the following for data sets with n observations, note the similarity to the definition of variance.

$$\text{Total Variation} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

We are interested in computing the percentage of the total variation of y that is explained by using the line of best fit to model the data; we call this percentage the **coefficient of determination** and denote it using the symbol R^2. To determine the coefficient of determination, we must be able to compute the explained variation in our model. Either the variation is explained or it is not explained. As such, we know that the total variation is equal to the sum of the unexplained variation and the explained variation. The disparity between predicted values and observed values is the source of the unexplained variation. At this point, we recognize that the SSE is the unexplained variation. Recall the meaning of the sum of the squared errors and think of the formula that would compute it in general.

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

With these ideas in hand we are ready to derive a computational formula for the coefficient of determination which measures the percent of variation in a data set that is accounted for by using the line of best fit to model the data.

$$R^2 = \frac{\text{Total Variation} - \text{SSE}}{\text{Total Variation}} = 1 - \frac{\text{SSE}}{\text{Total Variation}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Note that we dropped the indices on the summations in our formula to ease, but remember that for each of the summations, we are summing through all observations, 1 through n. The equivalency of the last computational formula is not obvious but can derived using the definitions of the line of best fit and the correlation coefficient, we encourage the reader to verify the simplification. In simplifying this last formulation, we develop the connection between the correlation coefficient and the coefficient of determination (a connection possibly already surmised given the symbolic representations); namely, that in squaring the correlation coefficient, we obtain the coefficient of determination.

Optional Derivation Connecting Correlation Coefficient and Coefficient of Determination for the Mathematically Inclined

$$\begin{aligned}
R^2 &= \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{\sum \left(r\frac{s_y}{s_x} x_i + \bar{y} - r\frac{s_y}{s_x}\bar{x} - \bar{y} \right)^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{\sum \left(r\frac{s_y}{s_x} x_i - r\frac{s_y}{s_x}\bar{x} \right)^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{\sum \left(r\frac{s_y}{s_x} (x_i - \bar{x}) \right)^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{\left(r\frac{s_y}{s_x} \right)^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{r^2 \frac{s_y^2}{s_x^2} (n-1) s_x^2}{(n-1) s_y^2} \\
&= \frac{r^2 \frac{\cancel{s_y^2}}{\cancel{s_x^2}} \cancel{(n-1)} \cancel{s_x^2}}{\cancel{(n-1)} \cancel{s_y^2}} = r^2
\end{aligned}$$

? Text Exercise 8.3.4

1. Using technology, confirm that the line of best fit matches the formula provided in the earlier text exercise when rounded to 4 decimal places. Using the slope and y-intercept of the line of best fit provided by technology without rounding, determine the SSE, the Total Variation,and R^2 by applying the formulas derived thus far in this section. Provide an interpretative sentence explaining the meaning of the coefficient of determination in light of the computed value. Do not round answers in intermediate steps.

Table 8.3.10: Values for variables x and y

Observation	x	y
1	7	5
2	2	4
3	6	6
4	3	3
5	1	2

Answer

Using the LINEST function in Excel, we confirm the accuracy of the previous text exercise rounded to 4 decimal places with a computed slope of $0.52238806\ldots$ and a y-intercept of $2.014925373\ldots$. We provide a table with values rounded to 6 decimals for checking.

Table 8.3.11: Computation of predicted values, squared errors, and variations

Observation	x	y	$\hat{y}$	$e_i^2 = (y-\hat{y})^2$	$(y_i - \bar{y})^2$
1	7	5	5.671642	$(-0.671642)^2 \approx 0.451113$	$(5-4)^2 = 1$
2	2	4	3.059702	$0.940299^2 \approx 0.884161$	$(4-4)^2 = 0$
3	6	6	5.149254	$0.850746^2 \approx 0.723769$	$(6-4)^2 = 4$
4	3	3	3.582090	$(-0.582090)^2 \approx 0.338828$	$(3-4)^2 = 1$
5	1	2	2.537313	$(-0.537313)^2 \approx 0.288716$	$(2-4)^2 = 4$

The sum of the squared errors is thus $\text{SSE} \approx 2.686567$, the smallest possible value for this particular data set. The Total Variation, the summation of the last column is 10. We can compute the coefficient of determination $R^2 \approx 1 - \frac{2.686567}{10} \approx 1 - 0.268657 \approx 0.731343$. Approximately 73.1% of the variation present in the y variable is accounted for using the linear function $\hat{y} = 0.522388x + 2.014925$. This indicates that the linear model fits the data to a certain degree, but there is a decent amount of random variation, error, or noise present.

2. Using the results of the previous part of this text exercise and technology, confirm that the square of the correlation coefficient is equal to the coefficient of determination. Compute the correlation coefficient using technology.

Answer

We computed in the previous part that $R^2 \approx 0.731343$. Using the CORREL function in excel, we compute that $r \approx 0.855186$ and note that $0.855186^2 \approx 0.731343$ which is the value that we computed for the coefficient of determination.

The coefficient of determination R^2 can be computed directly using the Excel function RSQ. The function takes two arrays of numbers, similar to the LINEST function, the first array consists of the known y-values (dependent variable) and the second array consists of the known x-values (independent variable).

Simple Linear Regression: Predictions and Interpretations

We have yet to conduct simple linear regression outside of a purely mathematical context. Having developed the concepts, we now address the application of these ideas and provide insight to their interpretations. Let us return to a data set that we have started to analyze, the ages of the bride and groom on their wedding day. Using a scatter plot of the data, we have already determined that a linear model would be appropriate. Let us determine the line of best fit and assess how well the model fits the data. In our previous considerations, we had the groom's age on the horizontal axis, the axis traditionally associated with the independent variable. For continuity of presentation, we will continue in this vein of thought. When evaluating the linear model we will be predicting the age of the bride based on the input of a groom's age. If we want to predict the age of a groom based on a particular age of a bride, we

will either have to solve for the age of the groom or conduct the linear regression analysis with the variables switched. Either option is fairly straightforward and will produce the same predictions.

Text Exercise 8.3.5

1. Letting the age of the bride on her wedding day be the dependent variable, determine the line of the best fit and the coefficient of determination for the data set. Explain the results of the linear regression in the context of the problem.

Table 8.3.12: Ages of bride and groom on wedding day

Married Couple	Groom's Age (years)	Bride's Age (years)
1	20	21
2	26	20
3	32	34
4	30	30
5	21	22
6	29	28
7	26	25
8	34	34
9	29	28
10	55	50
11	30	26
12	43	39
13	30	29
14	24	22
15	20	19

Answer

Using Excel, we have the line of best fit given by $\hat{y} = 0.878562x + 2.168364$ with a coefficient of determination equal to 0.938862. The computed coefficient of determination indicates that almost 94% of the variation in the age of a bride on her wedding day can be accounted by modeling the relationship between the ages of the bride and groom with the function $\hat{y} = 0.878562x + 2.168364$, where x is the age of the groom. This is a fairly high percentage which indicates that the model is a good fit. The positive slope indicates a positive association.

2. The y-intercept of a function is the value of the dependent variable when the independent variable is equal to 0. Within the context of our problem $x = 0$ corresponds to the groom's age being 0. The y-intercept is about 2.17, indicating that the bride would be just a little older than 2. Explain why, contextually, these considerations do not make any sense. What does this say about our model? What does this say about linear regression models in general?

Answer

Infants and toddlers do not get married. Adults get married. It is remiss to try to use the line of best fit to model the relationship where the relationship does not exist. There is a mathematical domain for our function and there is a contextual domain for our relation. If we are trying to understand the reality around us, the contextual domain must be at the forefront of our minds. We do not want to extend our model where the relationship ceases or beyond where our data permits us to engage. As such, we would not want to use our model for any ages less than 16 or 18 years of age for either the bride or the groom as those are the ages commonly set as the minimum ages for which marriage is legal. This does not say anything

negative about our model or models in general; we must be cognizant of when it is appropriate to use the models. Contextual clues are a big help. We will develop more nuance as we progress through this section.

3. Using the model constructed in part 1, predict the age of the bride when the groom is 27 years old.

Answer

We have that the line of best fit is given by $\hat{y} = 0.878562x + 2.168364$. We are being asked to predict the age of the bride when the groom is 27 years old. This is equivalent to evaluating the function when the x variable is 27. We predict the bride's age to be $0.878562 \cdot 27 + 2.168364 = 25.88955$ years old; the bride will be just shy of 26 years old, when the groom is 27 years old.

4. Using the model constructed in part 1, predict the age of the groom when the bride is 32 years old.

Answer

We have again that the line of best fit is given by $\hat{y} = 0.878562x + 2.168364$. We are asked to predict the age of the groom when the bride is 32 years old. This is not equivalent to evaluating the function when the x variable is 32 because the x variable corresponds to the age of the groom, not the bride. We predict the groom's age by solving for x when our linear equation equals 32.

$$\begin{aligned} 32 &= 0.878562x + 2.168364 \\ 32 - 2.168364 &= 0.878562x \\ \frac{32 - 2.168364}{0.878562} &= x \\ 33.955055 &\approx x \end{aligned}$$

We predict that the groom will be about just shy of 34 when the bride is 32 years old.

5. Using the model constructed in part 1, when does the model predict that the bride and groom will be exactly the same age? Does this seem like an appropriate use of the model?

Answer

For the last time, we have that the line of best fit is given by $\hat{y} = 0.878562x + 2.168364$. We want to find when the model predicts the two ages to be the same, i.e. $\hat{y} = x$. To do so, we replace $\hat{y}$ with x in the equation and solve.

$$\begin{aligned} x &= 0.878562x + 2.168364 \\ x - 0.878562x &= 2.168364 \\ x(1 - 0.878562) &= 2.168364 \\ x &= \frac{2.168364}{1 - 0.878562} \approx 17.855796 \end{aligned}$$

We predict the bride and groom to be the same age when they are both just shy of 18 on their wedding day.

In general, once a person is about 2 years of age, the primary focus is on the number of years. As such, our interest might be more of when the model predicts that both the bride and the groom would be in the same year of life. This would seem to be a more appropriate question given the context of the model; although, it is a much harder question to solve for ages of 17.5 and 17.8 would constitute solutions as well as what we just found.

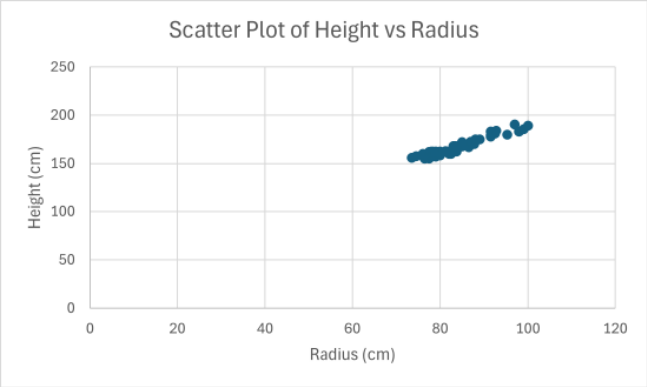
? Text Exercise 8.3.6

1. When ordering custom clothing or preparing to rent formal wear, many measurements are taken to ensure that the clothes fit well. Two common measurements are height and the length from the center of the back between the scapulae to the tips of the fingers when the arm is fully extended to the side. Let us refer to this latter measurement as an individual's radius.

We asked a random selection of 50 online elementary statistics students to obtain these measurements for themselves in centimeters and report their findings to analyze as a group. We provide the data in the attached Excel file. Examine the data to check that a linear model is appropriate. If not, explain. If so, find the line of best fit and coefficient of determination using the radius as the independent variable.

Answer

We first create a scatter plot to check if a linear relationship is reasonable. We provide two scatter plots with different scaling.



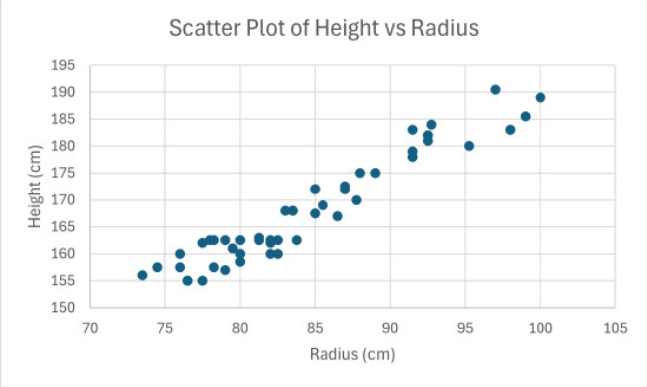


Figure 8.3.3 Scatter plots of height and radius using different scales and ranges

The scatter plot on the left includes the origin $(0, 0)$ while the other scatter plot does not. Both indicate a fairly linear relationship. So we proceed with a linear regression analysis. The coefficient of determination is $0.9212$with the linear model defined by $\hat{y} = 1.381x + 51.3125$.

In the previous text exercise, we determined the line of best fit and saw that the line fit fairly well. A little more than 92% of the variation in the height variable was attributed to the difference values of the radius variable through our linear model. We have a nice model to help us understand the relationship between the height and radius of individuals. The possible values of an individual's radius go beyond those collected in our sample. This is one of the reasons that we desired a model; so, that we could estimate values for points where we did not have any data collected. As such, we might be tempted to estimate the height of an individual with a radius of 40 centimeters.

? Text Exercise 8.3.7

Using the line of best fit found in the previous text exercise, estimate the height of an individual with a radius of 40 centimeters. Consider the validity of such an estimation.

Answer

We would estimate that individual's height to be $1.381 \cdot 40 + 51.3125 = 106.5525$centimeters. Both the radius and the height values are within the contextual domain of our variables, but can we use the model in such a way? The predicted

height is about three and a half feet tall; a rather short person. In reality, most likely a child. This begs the question: should we use data from a sample of elementary statistics students who are fully formed adults to make predictions about a child? Hopefully, at this stage in our development of statistics, we would be inclined to say no. We would not think that a sample of adults would be representative of children without some significant argumentation explaining why they are fundamentally the same. Our intuition would naturally be that the body structure of children is different than the body structure of adults. We do not want to overgeneralize our results beyond that which we have actually studied. In practice, we must consider both the contextual domain and the extent to which our sample is representative. In general, we do not want to utilize our model too far beyond the values seen in our collected data. Do you want to predict the height of an individual with a radius of 90 centimeters? Go right ahead! But, if you want to predict the height of an individual with a radius of $15,$ best go collect data from individuals around that size.

We conclude this section with one last interpretative guideline. The slope of a linear function describes the rate of change of the function. If the value of x increases by 1, the value of y changes in value equal to the slope. In the case of our last text exercise, when we increase the radius by one centimeter, the predicted y value increases by 1.381 centimeters. Are we to interpret this to indicate that if an individual had a radius of 75 centimeters and height of 150 centimeters and then grew to 76 centimeters, the individual's height would be 151.381 centimeters? Unfortunately, the answer is no. We built the model by using data from 50 individuals. The model predicts the typical relationship between the variables; it does not predict the individual change, nor does it predict the changes in a perfect way. We must temper ourselves from concluding more than we can. We can expect that as individuals increase in radius by 1 centimeter, the average gain in height is going to be close to 1.381 centimeters, but we cannot make such a claim on the individual level.

This is, in fact, a theme pertaining to the entirety of this textbook. Statistics seeks to understand trends in large groups, and it is almost always inappropriate to use information about a group to infer facts about an individual. If we say one group is shorter than another group on average, that does not necessarily mean that every individual in the first group is shorter than every individual in the second group. If we say that 80% of some group has a particular disease, that does not necessarily mean that each individual in that group has an 80% chance of obtaining that disease. If we say that a hypothesis or model predicts a group to have certain parameters, that says nothing about a specific individual in that group. Many issues in modern society arise from people misunderstanding this. People often use facts about a group to inform their thoughts about individuals (such as with stereotyping). People also often ignore facts about groups because of facts they know about an individual; for example, Bob smoked his whole life and lived to be 100. We hope that this text has helped the reader understand how to properly understand facts about groups and why such understanding can be useful.

Index

Detailed Licensing

Overview

Title: Elements of Statistics

Webpages: 66

All licenses found:

- Public Domain: 93.9% (62 pages)
- Undeclared: 6.1% (4 pages)

By Page

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Elements of Statistics

Webpages: 66

All licenses found:

- Public Domain: 93.9% (62 pages)
- Undeclared: 6.1% (4 pages)

By Page