MA336:STATISTICS

Fei Yi
Queensborough Community College

MA-336 Statistics

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

1: Introduction to Statistical Studies

- 1.1: Why It Matters- Types of Statistical Studies and Producing Data
- 1.2: Introduction to Types of Statistical Studies
- 1.3: Types of Statistical Studies (1 of 4)
- 1.4: Types of Statistical Studies (2 of 4)
- 1.5: Types of Statistical Studies (3 of 4)
- 1.6: Types of Statistical Studies (4 of 4)
- 1.7: Introduction to Sampling
- 1.8: Sampling (1 of 2)
- 1.9: Sampling (2 of 2)
- 1.10: Methods of Sampling
- 1.11: Introduction to Conducting Experiments
- 1.12: Conducting Experiments (1 of 2)
- 1.13: Conducting Experiments (2 of 2)
- 1.14: Putting It Together- Types of Statistical Studies and Producing Data

2: Descriptive Statistics

- 2.1: Organizing and Graphing Qualitative Data
- 2.2: Organizing and Graphing Quantitative Data
- 2.3: Stem-and-Leaf Displays
- 2.4: Measures of Central Tendency- Mean, Median and Mode
- 2.5: Measures of Position- Percentiles and Quartiles
- 2.6: Box Plots
- 2.7: Measures of Spread- Variance and Standard Deviation
- 2.8: Skewness and the Mean, Median, and Mode

3: Examining Relationships- Quantitative Data

- 3.1: Why It Matters- Examining Relationships- Quantitative Data
- 3.2: Linear Regression (4 of 4)
- 3.3: Introduction to Assessing the Fit of a Line
- 3.4: Assessing the Fit of a Line (1 of 4)
- 3.5: Assessing the Fit of a Line (2 of 4)
- 3.6: Assessing the Fit of a Line (3 of 4)
- 3.7: Assessing the Fit of a Line (4 of 4)
- 3.8: Putting It Together- Examining Relationships- Quantitative Data
- 3.9: StatTutor- Academic Performance
- 3.10: Assignment- Scatterplot
- 3.11: Assignment- Linear Relationships
- 3.12: Introduction to Scatterplots
- 3.13: Assignment- Linear Regression
- 3.14: Scatterplots (1 of 5)
- 3.15: Scatterplots (2 of 5)
- 3.16: Scatterplots (3 of 5)
- 3.17: Scatterplots (4 of 5)
- 3.18: Scatterplots (5 of 5)

- 3.19: Introduction to Linear Relationships
- 3.20: Linear Relationships (1 of 4)
- 3.21: Linear Relationships (2 of 4)
- 3.22: Linear Relationships (3 of 4)
- 3.23: Linear Relationships (4 of 4)
- 3.24: Introduction to Association vs Causation
- 3.25: Causation and Lurking Variables (1 of 2)
- 3.26: Causation and Lurking Variables (2 of 2)
- 3.27: Introduction to Linear Regression
- 3.28: Linear Regression (1 of 4)
- 3.29: Linear Regression (2 of 4)
- 3.30: Linear Regression (3 of 4)

4: Relationships in Categorical Data with Intro to Probability

- 4.1: Why It Matters- Relationships in Categorical Data with Intro to Probability
- 4.2: Introduction to Two-Way Tables
- 4.3: Two-Way Tables (1 of 5)
- 4.4: Two-Way Tables (2 of 5)
- 4.5: Two-Way Tables (3 of 5)
- 4.6: Two-Way Tables (4 of 5)
- 4.7: Two-Way Tables (5 of 5)
- 4.8: Putting It Together- Relationships in Categorical Data with Intro to Probability

5: Basic Concepts of Probability

- 5.1: Sample Spaces, Events, and Their Probabilities
- 5.2: Complements, Intersections, and Unions
- 5.3: Conditional Probability and Independent Events
- 5.E: Basic Concepts of Probability (Exercises)

6: Discrete Random Variables

- 6.1: Random Variables
- 6.2: Probability Distributions for Discrete Random Variables
- 6.3: The Binomial Distribution
- 6.E: Discrete Random Variables (Exercises)

7: Continuous Random Variables

- 7.1: Continuous Random Variables
- 7.2: The Standard Normal Distribution
- 7.3: Probability Computations for General Normal Random Variables
- 7.4: Areas of Tails of Distributions
- 7.E: Continuous Random Variables (Exercises)

8: Sampling Distributions

- 8.1: The Mean and Standard Deviation of the Sample Mean
- 8.2: The Sampling Distribution of the Sample Mean
- 8.3: The Sample Proportion
- 8.4: Using the Central Limit Theorem
  - 8.4E: Using the Central Limit Theorem (Exercises)
- 8.E: Sampling Distributions (Exercises)

9: Confidence Intervals

- 9.1: Prelude to Confidence Intervals
- 9.2: A Single Population Mean using the Normal Distribution
- 9.3: A Single Population Mean using the Student t-Distribution
- 9.4: A Population Proportion
- 9.5: Confidence Interval - Home Costs (Worksheet)
- 9.6: Confidence Interval -Place of Birth (Worksheet)
- 9.7: Confidence Interval -Women's Heights (Worksheet)
- 9.E: Confidence Intervals (Exercises)
- 9.S: Confidence Intervals (Summary)

10: Hypothesis Testing with One Sample

- 10.1: Prelude to Hypothesis Testing
- 10.2: Null and Alternative Hypotheses
- 10.3: Outcomes and the Type I and Type II Errors
- 10.4: Distribution Needed for Hypothesis Testing
- 10.5: Rare Events, the Sample, Decision and Conclusion
- 10.6: Additional Information and Full Hypothesis Test Examples
- 10.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet)
- 10.E: Hypothesis Testing with One Sample (Exercises)

Index

Glossary

Detailed Licensing

Licensing

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

CHAPTER OVERVIEW

1: Introduction to Statistical Studies

1.1: Why It Matters- Types of Statistical Studies and Producing Data
1.2: Introduction to Types of Statistical Studies
1.3: Types of Statistical Studies (1 of 4)
1.4: Types of Statistical Studies (2 of 4)
1.5: Types of Statistical Studies (3 of 4)
1.6: Types of Statistical Studies (4 of 4)
1.7: Introduction to Sampling
1.8: Sampling (1 of 2)
1.9: Sampling (2 of 2)
1.10: Methods of Sampling
1.11: Introduction to Conducting Experiments
1.12: Conducting Experiments (1 of 2)
1.13: Conducting Experiments (2 of 2)
1.14: Putting It Together- Types of Statistical Studies and Producing Data

---

1.1: Why It Matters- Types of Statistical Studies and Producing Data

Why learn about the various types of statistical studies and how data is produced?

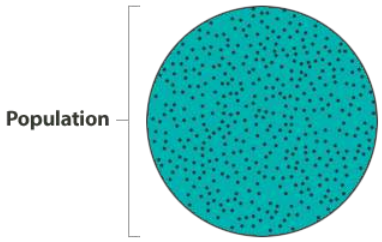
We organized this course around the Big Picture of Statistics. As we learn new material, we will always look at how these new ideas relate to the Big Picture. In this way the Big Picture is a diagram that will help us organize and understand the material we will learn throughout the course.

The Big Picture summarizes the steps in a statistical investigation.

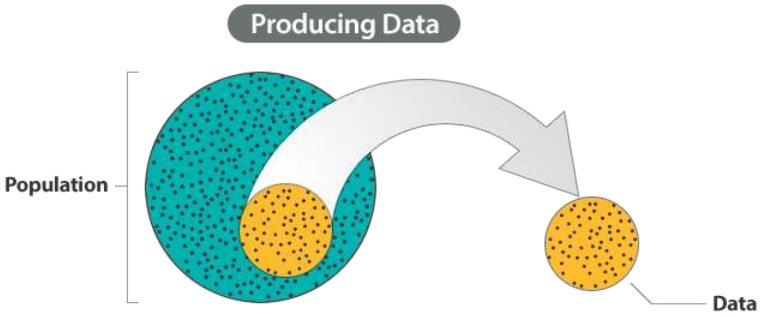
We begin a statistical investigation with a research question. The research question is frequently something we want to know about a **population**. The population can be people or other things, such as animals or objects. For example, we might want to know the answer to questions such as:

- What percentage of U.S. adults supports the death penalty? (Population: U.S. adults)
- Do cell phones affect bees? (Population: bees)
- Do cars get better gas mileage with a new gasoline additive? (Population: cars)

The population is the entire group that we want to know something about:



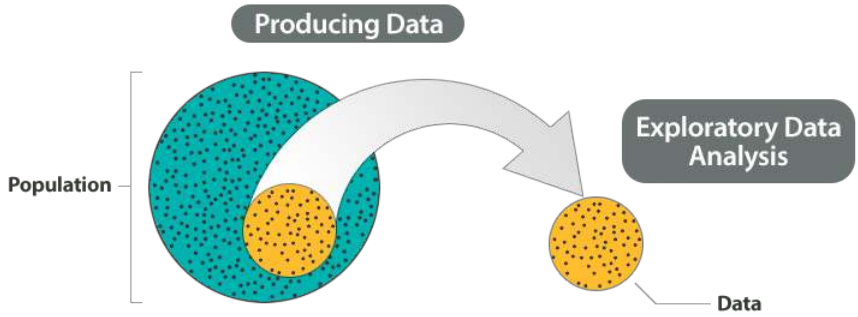
In most cases, the population is a large group. Often, the population is so large that we cannot collect information from every individual in the population. So we select a **sample** from the population. Then we collect data from this sample. This is the first step in the statistical investigation. We call this step **producing data**.



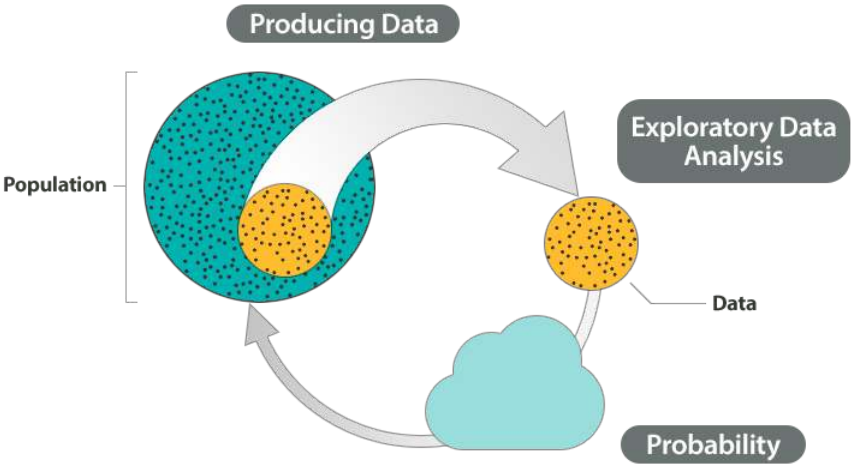
Of course, we need a sample that represents the population well. This involves careful planning but also involves chance. For example, if our goal is to determine the percentage of U.S. adults who favor the death penalty, we do not want our sample to contain only Democrats or only Republicans. So we can give everyone the same opportunity to be in the sample, but we will let chance select the sample.

At this step of the investigation we also carefully define what kind of information we plan to gather. Then we collect the data.

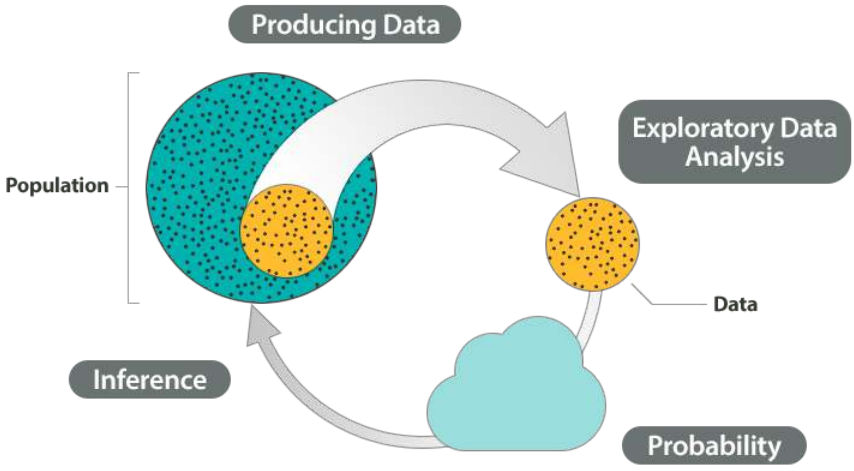
Data is often a long list of information. To make sense of the data, we explore it and summarize it using graphs and different numerical measures, such as percentages or averages. We call this step **exploratory data analysis**.



Remember, our goal is to answer a question about a population based on a sample. Of course, samples will vary due to chance, and we will need to answer our question in spite of this variability. So we need to understand how sample results will vary and how sample results relate to the population as a whole when chance is involved. This is where **probability** comes in.



Probability is the "machinery" behind the last step in the process called **inference**. We infer something about a population based on a sample. This inference is the conclusion we reach from our sample data that answers our original question about the population.



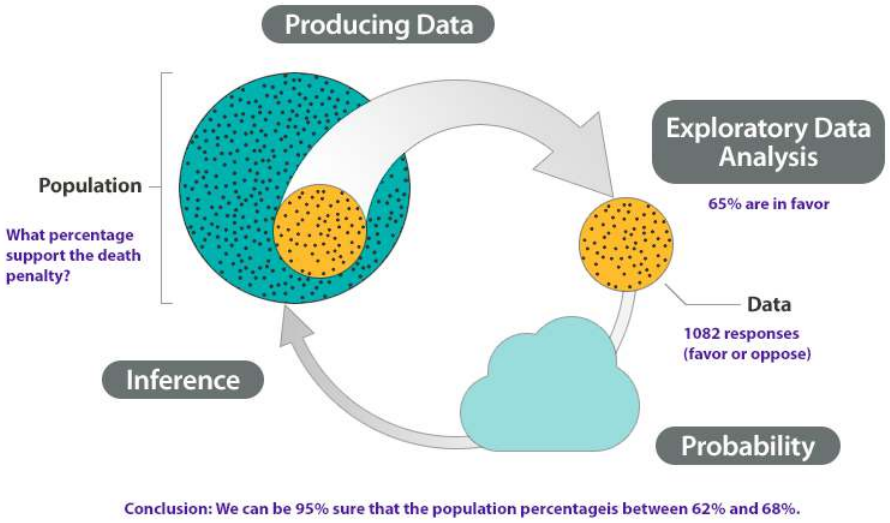
Example – The big picture of statistics

At the end of April 2005, ABC News and the Washington Post conducted a poll to determine the percentage of U.S. adults who support the death penalty.

Research question: What percentage of U.S. adults support the death penalty?

Steps in the statistical investigation:

1. **Produce Data**: *Determine what to measure, then collect the data.*
   The poll selected 1,082 U.S. adults at random. Each adult answered this question: "Do you favor or oppose the death penalty for a person convicted of murder?"
2. **Explore the Data**: *Analyze and summarize the data.*
   In the sample, 65% favored the death penalty.
3. **Draw a Conclusion**: *Use the data, probability, and statistical inference to draw a conclusion about the population.*
   Our goal is to determine the percentage of the U.S. adult population that supports the death penalty. We know that different samples will give different results. What are the chances that our sample reflects the opinions of the population within 3%? Probability describes the likelihood that our sample is this accurate. So we can say with 95% confidence that between 62% and 68% of the population favor the death penalty.



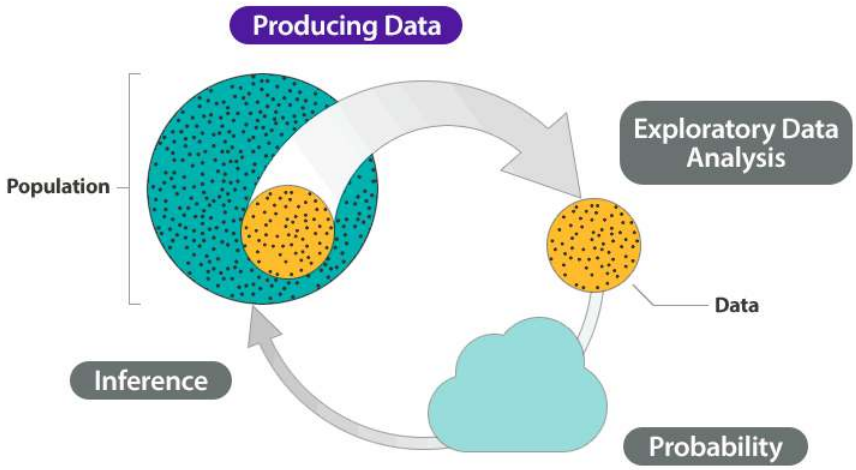
Let's Summarize

A statistical investigation with a research question. Then the investigation proceeds with the following steps:

- Produce Data: Determine what to measure, then collect the data.
- Explore the Data: Analyze and summarize the data (also called *exploratory data analysis*).
- Draw a Conclusion: Use the data, probability, and statistical inference to draw a conclusion about the population.

Types of Statistical Studies and Producing Data

In this first module, we focus on the *produce data* step in a statistical investigation. We discuss two types of statistical investigations: the observational study and the experiment. Each type of investigation involves a different approach to collecting data. We will also see that our approach to collecting data determines what we can conclude from the data.



Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

1.2: Introduction to Types of Statistical Studies

What you'll learn to do: Describe various types of statistical studies and the types of conclusions that are appropriate.

In statistical studies, the type of study design used and the details of the design are important in determining what kind of conclusions we may draw from the results. In particular, simply observing an association between two variables – say, smoking and cancer – does not guarantee that one variable causes the other. In this section, we will explore how the details of a study design play a crucial role in determining our ability to establish evidence of causation.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

1.3: Types of Statistical Studies (1 of 4)

Learning Objectives

- From a research question, determine the goal of a statistical study.

Before we begin our discussion of the types of statistical studies, we look closely at the types of research questions used in statistical studies.

Research Questions about a Population

Recall that a *population* is the entire group of individuals or objects that we want to study. Usually, it is not possible to study the whole population, so we collect data from a part of the population, called a *sample*. We use the sample to draw conclusions about the population.

For example, suppose our research question is "What is the average amount of money spent on textbooks per semester by full-time students at Seattle Central?" We cannot interview every full-time student at Seattle Central because would take too much time and cost too much money. We therefore carefully select a sample of full-time students at Seattle Central to represent the population of all full-time students at that college. Then we collect data from the sample to estimate the average amount spent on textbooks.

This example illustrates how the research question guides the investigation. A well-stated research question contains information about:

- The population (full-time students at Seattle Central).
- The information we will collect from each individual in the sample. We also call this the **variable**. The variable is what we plan to measure (amount of money spent on textbooks per semester).
- A numerical characteristic about the population related to this variable (the *average* amount of money spent on textbooks per semester).

Here are some common types of research questions about a population:

Type of Research Question	Examples
Make an estimate about the population (often an estimate about an *average* value or a *proportion* with a given characteristic)	What is the *average* number of hours that community college students work each week?
	What *proportion* of all U.S. college students are enrolled at a community college?
Test a claim about the population (often a claim about an *average* value or a *proportion* with a given characteristic)	Is the *average* course load for a community college student greater than 12 units?
	Do the *majority* of community college students qualify for federal student loans?
Compare two populations (often a comparison of population averages or proportions with a given characteristic)	In community colleges, do female students have a *higher* GPA than male students?
	Are college athletes *more* likely than nonathletes to receive academic advising?
Investigate a relationship between two variables in the population	Is there a *relationship* between the number of hours high school students spend each week on Facebook and their GPA?
	Is academic counseling *associated* with quicker completion of a college degree?

Try It

https://assessments.lumenlearning.co...sessments/3820

Try It

https://assessments.lumenlearning.co...sessments/3821

Research Questions about Cause and Effect

A research question that focuses on a **cause-and-effect** relationship is common in disciplines that use experiments, such as medicine or psychology. These types of questions ask how one variable responds as another variable is manipulated. These types of questions involve two variables. Here are some examples:

- Does cell phone usage increase the risk of developing a brain tumor?
- Does drinking red wine lower the risk of a heart attack?
- Does playing violent video games increase aggressive behavior?
- Does sex education lower the incidence of teen pregnancy?

To provide convincing evidence of a cause-and-effect relationship, the researcher designs an experiment. We discuss experiments in “Collecting Data – Conducting an Experiment.”

Note: In the previous section, *Research Questions about a Population*, we included examples of questions about the relationship between two variables in a population. But in these types of questions, we used words like *associated*, *correlated*, *linked to*, and *connected*. These words do not imply a cause-and-effect relationship between the variables. We can investigate these types of questions without conducting an experiment – an observational study will do. We study observational studies in “Collecting Data – Sampling.”

Try It

https://assessments.lumenlearning.co...sessments/3822

Contributors and Attributions

1.4: Types of Statistical Studies (2 of 4)

Learning Objectives

- Determine if a study is an experiment or an observational study.
- From a description of a statistical study, determine the goal of the study.

In general, there are two types of statistical studies: observational studies and experiments.

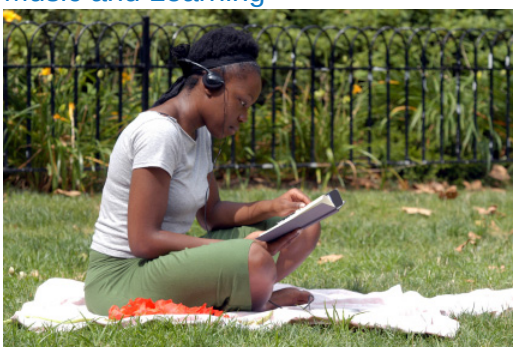
An **observational study** observes individuals and measures variables of interest. The main purpose of an observational study is to describe a group of individuals or to investigate an association between two variables. We can answer questions about a population with an observational study. We can also investigate a relationship between two variables. But in an observational study, researchers do not attempt to manipulate one variable to cause an effect in another variable. For this reason, an observational study does not provide convincing evidence of a cause-and-effect relationship.

An **experiment** intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables. But the experiment has to be well-designed to provide convincing evidence of a cause-and-effect relationship. We study experiment design in the section "Collecting Data – Conducting an Experiment."

For now, our goal is to distinguish between these two types of studies. We focus on the connection between the research question, the type of study, and the kinds of conclusions we can make.

Example

Music and Learning



Many students listen to music while studying. Does listening to music improve learning? Students in a statistics class decide to investigate this question. They write more specific research questions related to the topic of music and learning. Then they design the following three studies:

Study 1

Specific research questions: *Do the majority of college students listen to music while they study? Do the majority of college students believe that listening to music improves their learning?*

To investigate these questions, the statistics students conduct a survey in their other classes. They ask these two questions:

- Do you listen to music while you study?
- Do you think listening to music improves your concentration and memory?

This is an observational study designed to answer two questions about a population of college students. Each question contains a claim about the population of college students. We can use data from this study to see if these claims are true. But data from this study cannot provide evidence of a cause-and-effect relationship between listening to music while studying and improvements in learning.

Study 2

Specific research question: *When we compare students who study with music to students who study in a quiet environment, which group gives higher ratings for understanding what they studied?*

To investigate this question, the instructor divides the class into two groups: (1) those who listen to music when they study and (2) those who do not listen to music when they study. The students keep a journal for a week. Each time they study, they record the following information:

- Length of study session (in minutes)
- A rating of how well they understood what they studied, on a scale of 1–10: 1 = no understanding, 10 = excellent understanding.

This investigation is also an observational study. It compares two populations: (1) college students who listen to music when studying and (2) college students who do not listen to music when studying. We can also view this as an observational study of one population (college students) that investigates the relationship between two variables: *listening to music while studying* and *perceived understanding of material studied*. From this study, we might learn something interesting about the connection between college students' study habits and their perception of their learning. But since this is an observational study, data from this study cannot provide evidence of a cause-and-effect relationship between listening to music while studying and improvements in learning.

Study 3

Specific research question: *Does listening to music improve students' ability to quickly identify information?*

To investigate this question, the instructor uses word-search puzzles. She divides the class into two groups. Students on one side of the room do a word puzzle for 3 minutes while listening to music on an iPod. Students on the other side of the room do a word puzzle for 3 minutes without music. The instructor calculates the average number of words found by each group.

This study is an experiment. The instructor manipulates music to investigate the impact on puzzle completion. Data from this study can provide evidence of a cause-and-effect relationship between listening to music while studying and improvements in learning. But the improvement in learning is more narrowly defined as the ability to quickly identify information, such as words in a puzzle.

Try It

https://assessments.lumenlearning.co...sessments/3401

https://assessments.lumenlearning.co...sessments/3402

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: oli.cmu.edu. **License**: *CC BY: Attribution*

1.5: Types of Statistical Studies (3 of 4)

Learning Objectives

- Based on the study design, determine what types of conclusions are appropriate.

We now focus more closely on studies that investigate a relationship between two variables. In these types of studies, one variable is the **explanatory variable**, and the other is the **response variable**. To establish a cause-and-effect relationship, we want to make sure the explanatory variable is the only thing that impacts the response variable. We therefore get rid of all other factors that might affect the response. Then we manipulate the explanatory variable. Our goal is to see if it really does affect the response.

In an observational study, researchers may take steps to reduce the influence of these other factors on the response. But it is difficult in an observational study to get rid of all the factors that may have an influence. In addition, the researchers do not manipulate the explanatory variable to see if it affects the response. They just collect data and look for an association between the two variables. For these reasons, observational studies do not give convincing evidence of a cause-and-effect relationship.

In an experiment, researchers use a variety of techniques to eliminate the influence of these other factors. Then they manipulate the explanatory variable to see if it affects the response. For this reason, experiments give the strongest evidence for a cause-and-effect relationship.

Example

Hormone Replacement Therapy

When women go through menopause, the production of hormones in their bodies changes. The hormonal changes can cause a variety of symptoms that may be reduced by hormone replacement therapy. In the 1980s, hormone replacement therapy was common in the United States.

In the early 1990s, observational studies suggested that hormone replacement therapy had additional benefits, including a reduction in the risk of heart disease. In these observational studies, researchers compared women who took hormones to those who did not take hormones. Health records showed that women taking hormones after menopause had a lower incidence of heart disease. But women who take hormones are different from other women. They tend to be richer and more educated, to have better nutrition, and to visit the doctor more frequently. These women have many habits and advantages that contribute to good health, so it is not surprising that they have fewer heart attacks. But can we conclude from these studies that the hormones caused the reduction in heart attacks? No. The results are *confounded* by the influence of these other factors.

In 2002, the Women's Health Initiative sponsored a large-scale, well-designed experiment to study the health implications of hormone replacement therapy. In this experiment, researchers randomly assigned over 16,000 women to one of two treatments. One group took hormones. The other group took a **placebo**. A placebo is a pill with no active ingredients that looks like the hormone pill. The experiment was **double-blind**. *Blind* means that women did not know if they were receiving hormones or the placebo. *Double-blind* means that the information was coded, so researchers administering the pills did not know which treatment the women received. After 5 years, the group taking hormones had a *higher* incidence of heart disease and breast cancer. This is exactly the opposite result from the result found in the observational studies! In fact, the differences were so significant that the researchers ended the experiment early. The National Institutes of Health declared that the observational studies were wrong. Hormone replacement therapy to treat menopausal symptoms is now rarely used.

What's the Main Point?

An observational study can provide evidence of a link or an association between two variables. But other factors may also influence the results. These other factors are called *confounding variables*. The influence of confounding variables on the response variable is one of the reasons that an observational study gives weak, and potentially misleading, evidence of a cause-and-effect relationship. A well-designed experiment takes steps to eliminate the effects of confounding variables, including random assignment of people to treatment groups, use of a placebo, or blind conditions. Using these precautions, a well-designed experiment provides convincing evidence of cause-and-effect.

Try It

https://assessments.lumenlearning.co...sessments/3404

https://assessments.lumenlearning.co...sessments/3405

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

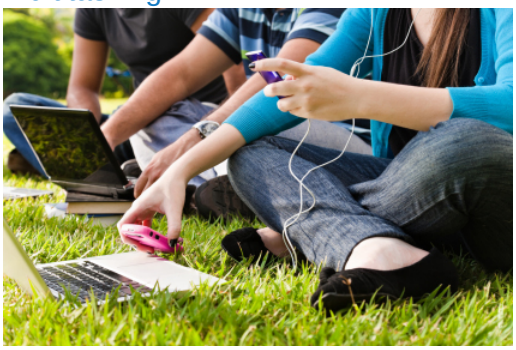
1.6: Types of Statistical Studies (4 of 4)

Learning Objectives

- Based on the study design, determine what types of conclusions are appropriate.

Example

Multitasking



Do you constantly text-message while in class? Do you jump from one website to another while doing homework? If so, then you are a high-tech multitasker. In a study of high-tech multitasking at Stanford University, researchers put 100 students into two groups: those who regularly do a lot of media multitasking and those who don't. The two groups performed a series of three tasks:

(1) A task to measure the ability to pay attention:

Students view two images of red and blue rectangles flashed one after the other on a computer screen. They try to tell if the red rectangles are in a different position in the second frame.

(2) A task to measure control of memory:

Students view a sequence of letters flashed onto a computer screen, then recall which letters occurred more than once.

(3) A task to measure the ability to switch from one job to another:

Students view numbers and letters together with the instructions to pay attention to the numbers, then recall if the numbers were even or odd. Then the instructions switch. Students are to pay attention to the letters and recall if the letters were vowels or consonants.

On every task, the multitaskers did worse than the non-multitaskers.

The researchers concluded that "people who are regularly bombarded with several streams of electronic information do not pay attention, control their memory, or switch from one job to another as well as those who prefer to complete one task at a time" (as reported in *Stanford News* in 2009).

"When they're [high-tech multitaskers] in situations where there are multiple sources of information coming from the external world or emerging out of memory, they're not able to filter out what's not relevant to their current goal," said Wagner, an associate professor of psychology at Stanford. "That failure to filter means they're slowed down by that irrelevant information."

Try It

https://assessments.lumenlearning.co...sessments/3403

In general, we should not make cause-and-effect statements from observational studies, but in reality, researchers do it all the time. This does not mean that researchers are drawing incorrect conclusions from observational studies. Instead, they have developed techniques that go a long way toward decreasing the impact of confounding variables. These techniques are beyond the scope of this course, but we briefly discuss a simplified example to illustrate the idea.

Example

Smoking and Cancer

Consider this excerpt from the National Cancer Institute website:

> *Smoking is a leading cause of cancer and of death from cancer. Millions of Americans have health problems caused by smoking. Cigarette smoking and exposure to tobacco smoke cause an estimated average of 438,000 premature deaths each year in the United States.*

Notice that the National Cancer Institute clearly states a cause-and-effect relationship between smoking and cancer. Now let's think about the evidence that is required to establish this causal link. Researchers would need to conduct experiments similar to the hormone replacement therapy experiments done by the Women's Health Initiative. Such experiments would be very difficult to do. The researchers cannot manipulate the smoking variable. Doing so would require them to randomly assign people to smoke or to abstain from smoking their whole life. Obviously, this is impossible. So how can we say that smoking *causes* cancer?

In practice, researchers approach this challenge in a variety of ways. They may use advanced techniques for making *statistical adjustments* within an observational study to control the effects of confounding variables that could influence the results. A simple example is the cell phone and brain cancer study.

> *In this observational study, researchers identified a group of 469 people with brain cancer. They paired each person who had brain cancer with a person of the same sex, of similar age, and of the same race who did not have brain cancer. Then they compared the cell phone use for each pair of people. This matching attempts to control the confounding effects of sex, age, and race on the response variable, cancer. With these adjustments, the study will provide stronger evidence for (or against) a casual link.*

However, even with such adjustments, we should be cautious about using evidence from an observational study to establish a cause-and-effect relationship. Researchers used these types of adjustments in the observational studies with hormone replacement therapy. We saw in that research that the results were still misleading when compared to those of an experiment.

So how can the National Cancer Institute state as a fact that smoking causes cancer?

They used other nonstatistical guidelines to build evidence for a cause-and-effect relationship from observational studies. In this approach, researchers review a large number of observational studies with criteria that, if met, provide stronger evidence of a possible cause-and-effect relationship. Here are some simplified examples of the criteria they use:

(1) There is a reasonable explanation for how one variable might cause the other.

- For example, experiments with rats show that chemicals found in cigarettes cause cancer in rats. It is therefor reasonable to infer that these same chemicals may cause cancer in humans.
- Consider these experiments together with the observational studies showing the association between smoking and cancer in humans. We now have more convincing evidence of a possible cause-and-effect relationship between smoking and cancer in humans.

(2) The observational studies vary in design so that factors that confound one study are not present in another.

- For example, one observational study shows an association between smoking and lung cancer, but the people in the study all live in a large city. Air pollution in a large city may contribute to the lung cancer, so we cannot be sure that smoking is the cause of cancer in this study.
- Another observational study looks only at nonsmokers. This study shows no difference in lung cancer rates for nonsmokers living in rural areas compared to nonsmokers living in cities.

- Consider these two studies together. The second study suggests that air pollution does not contribute to lung cancer, so we now have more convincing evidence that smoking (not air pollution) is the cause of higher cancer rates in the first study.

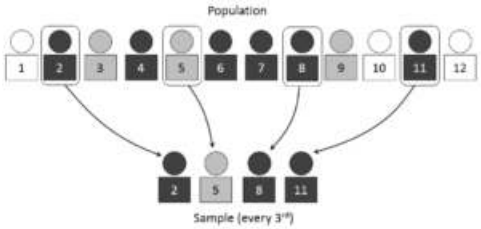
Let's Summarize

- There are four steps in a statistical investigation:
  - Ask a question that can be answered by collecting data.
  - Decide what to measure, and then collect data.
  - Summarize and analyze.
  - Draw a conclusion, and communicate the results.
- There are two types of statistical research questions:
  - Questions about a population
  - Questions about cause-and-effect
- To answer a question about a population, we select a sample and conduct an observational study. To answer a question about cause-and-effect we conduct an experiment.
- There are two types of statistical studies:
  - Observational studies: An *observational study* observes individuals and measures variables of interest. We conduct observational studies to investigate questions about a population or about an association between two variables. An observational study alone does not provide convincing evidence of a cause-and-effect relationship.
  - Experiments: An *experiment* intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.
- In statistics, a *variable* is information we gather about individuals or objects.
- When we investigate a relationship between two variables, we identify an *explanatory* variable and a *response* variable. To establish a cause-and-effect relationship, we want to make sure the explanatory variable is the only thing that impacts the response variable. Other factors, however, may also influence the response. These other factors are called *confounding* variables.
- The influence of confounding variables on the response variable is one of the reasons that an observational study gives weak, and potentially misleading, evidence of a cause-and-effect relationship. A well-designed experiment takes steps to eliminate the effects of confounding variables, such as random assignment of people to treatment groups, use of a placebo, and blind conditions. For this reason, a well-designed experiment provides convincing evidence of cause-and-effect.

Contributors and Attributions

1.7: Introduction to Sampling

What you'll learn to do: For an observational study, critique the sampling plan. Recognize implications and limitations of the plan.



Statistics seeks to use information about variables or relationships from a statistical study (sample) to draw conclusions about what is true for the entire population from which the sample was chosen. For this process to work reliably, it is essential that the sample be truly representative of the larger population. In this section, we will look at how we can create a sampling plan so that the sampling is carried out in such a way that the sample really does represent the population of interest.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: oli.cmu.edu. **License**: *CC BY: Attribution*

1.8: Sampling (1 of 2)

Learning Objectives

- For an observational study, critique the sampling plan. Recognize implications and limitations of the plan.

We now focus on observational studies and how to collect reliable and accurate data for an observational study.

We know that an observational study can answer questions about a population. But populations are generally large groups, so we cannot gather data from every individual in the population. Instead, we select a sample and gather data from the sample. We use the data from the sample to make statements about the population.

Here are two examples:

- A political scientist wants to know what percentage of college students consider themselves conservatives. The population is college students. It would be too time consuming and expensive to poll every college student, so the political scientist selects a sample of college students. Of course, the sample must be carefully selected to represent the political perspectives that are present in the population.
- A government agency plans to test airbags from Honda to determine if the airbags work properly. Testing an airbag means it has to be inflated and punctured, which ruins the airbag, so the researchers certainly cannot test every airbag. Instead, they test a sample of airbags and draw a conclusion about the quality of airbags from Honda.

Important Point

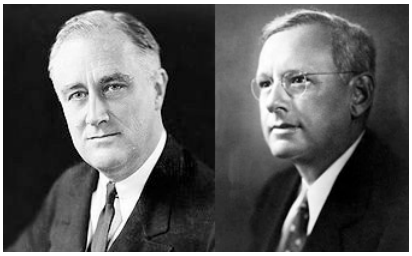
Our goal is to use a sample to make valid conclusions about a population. Therefore, the *sample must be representative of the population*. A representative sample is a subset of the population that reflects the characteristics of the population.

A **sampling plan** describes exactly how we will choose the sample. A sampling plan is **biased** if it systematically favors certain outcomes.

In our discussion of sampling plans, we focus on surveys. The next example is a famous one that illustrates how biased sampling in a survey leads to misleading conclusions about the population.

Example

The 1936 Presidential Election



In 1936, Democrat Franklin Roosevelt and Republican Alf Landon were running for president. Before the election, the magazine *Literary Digest* sent a survey to 10 million Americans to determine how they would vote. More than 2 million people responded to the poll; 60% supported Landon. The magazine published the findings and predicted that Landon would win the election. However, Roosevelt defeated Landon in one of the largest landslide presidential elections ever.

What happened?

The magazine used a biased sampling plan. They selected the sample using magazine subscriptions, lists of registered car owners, and telephone directories. The sample was not representative of the American public. In the 1930s, Democrats were much less likely to own a car or have a telephone. The sample therefore systematically *underrepresented* Democrats. The poll results did not represent the way people in the general population voted.

Before we discuss a method for avoiding bias, let’s look at some examples of common survey plans that produce unreliable and potentially biased results.

Example

How to Sample Badly

Online polls: The American Family Association (AFA) is a conservative Christian group that opposes same-sex marriage. In 2004, the AFA began a campaign in support of a constitutional amendment to define marriage as strictly between a man and a woman. The group posted a poll on its website asking AFA members to voice their opinion about same-sex marriage. The AFA planned to forward the results to Congress as evidence of America's opposition to same-sex marriage. Almost 850,000 people responded to the poll. In the poll, 60% *favored* legalizing same-sex marriage.

What happened? Against the wishes of the AFA, the link to the poll appeared in blogs, social-networking sites, and a variety of email lists connected to gay/lesbian/bisexual groups. The AFA claimed that gay rights groups had *skewed* its poll. Of course, the results of the poll would have been skewed in the other direction had only AFA members been allowed to participate.

This is an example of a **voluntary response sample**. The people in a voluntary response sample are self-selected, not chosen. For this reason, a voluntary response sample is biased because only people with strong opinions make the effort to participate.

Mall surveys: Have you ever noticed someone surveying people at a mall? People shopping at a mall are more likely to be teenagers, retired people, or people who have more money than the typical American. In addition, unless interviewers are carefully trained, they tend to interview people with whom they are comfortable talking. For these reasons, mall surveys frequently *overrepresent* the opinions of white middle-class or retired people. Mall surveys are an example of a **convenience sample**.

Example

How to Eliminate Bias in Sampling

In a voluntary response sample, people choose whether to respond. In a convenience sample, the interviewer chooses who will be part of the sample. In both cases, personal choice produces a biased sample. **Random sampling** is the best way to eliminate bias. Collecting a random sample is like pulling names from a hat (assuming every individual in the population has a name in the hat!). In a **simple random sample** everyone in the population has an equal chance of being chosen.

Reputable polling firms use techniques that are more complicated than pulling names out of a hat. But the goal is the same: eliminate bias by using random chance to decide who is in the sample.

Random samples will eliminate bias, even bias that may be hidden or unknown. The next three activities will reveal a bias that most of us have but don't know that we have! We will see how random sampling avoids this bias.

Random Samples

Instructions: Use the simulation below for this activity. You will see 60 circles. This is the "population." *Our goal is to estimate the average diameter of these 60 circles by choosing a sample.*

1. Choose a sample of five circles that look representative of the population of all 60 circles. Mark your five circles by clicking on each of them. They will turn orange. Record the average diameter for the five circles. (Make sure you have five orange circles before you record the average diameter.)
2. Reset the simulation.
3. Choose another five circles and record the average diameter for this sample of circles. You can reuse a circle, but the sample should not have all the same circles. You now have the averages for two samples.
4. Reset and repeat for a total of 10 samples. Record the average diameter for each sample.

Click here to open this simulation in its own window.

A link to an interactive elements can be found at the bottom of this page.

Try It

https://assessments.lumenlearning.co...sessments/3823

https://assessments.lumenlearning.co...sessments/3824

Now we estimate the average diameter of the 60 circles using *random* samples.

Instructions: Use the simulation below for this activity. You will again see the same 60 circles. As before, this is the "population." *Our goal is to estimate the average diameter of these 25 circles by choosing a random sample.*

1. Click on the “Generate sample” button to get a random sample of five circles by clicking on the random sample button. The simulation randomly chooses five circles. Record the average diameter for the random sample.
2. Reset the simulation using the reset button.
3. Click on the “Generate sample” button to get another random sample. Record the average diameter for this random sample. You now have the averages for two samples.
4. Reset and repeat for a total of 10 samples. Record the average diameter for each sample.

Click here to open this simulation in its own window.

A link to an interactive elements can be found at the bottom of this page.

Try It

https://assessments.lumenlearning.co...sessments/3825

https://assessments.lumenlearning.co...sessments/3826

Try It

https://assessments.lumenlearning.co...sessments/3827

Comment

Random selection also guarantees that the sample results do not change haphazardly from sample to sample. When we use random selection, the variability we see in sample results is due to chance. The results obey the mathematical laws of probability. We looked at this idea briefly in the Big Picture of Statistics. Probability is the machinery for drawing conclusions about a population on the basis of samples. To use this machinery, the sample must be chosen by random chance.

Try It

https://assessments.lumenlearning.co...sessments/3406

Try It

https://assessments.lumenlearning.co...sessments/3407

Contributors and Attributions

1.9: Sampling (2 of 2)

Learning Objectives

- For an observational study, critique the sampling plan. Recognize implications and limitations of the plan.

Let's briefly summarize the main points about sampling:

- We draw a conclusion about the population on the basis of the sample.
- To draw a valid conclusion, the sample must be representative of the population. A representative sample is a subset of the population that reflects the characteristics of the population.
- A sample is biased if it systematically favors a certain outcome.
- Random selection eliminates bias.

We have not mentioned the size of the sample. Are larger samples more accurate? Well, the answer is yes and no.

Recall the 1936 presidential election. A sample of over 2 million people did not correctly identify the winner of the election. Two million people is a huge sample, yet the results were completely wrong. So a large sample does not guarantee reliable results.

However, if the samples are randomly selected, then size does matter. We see this in the next example.

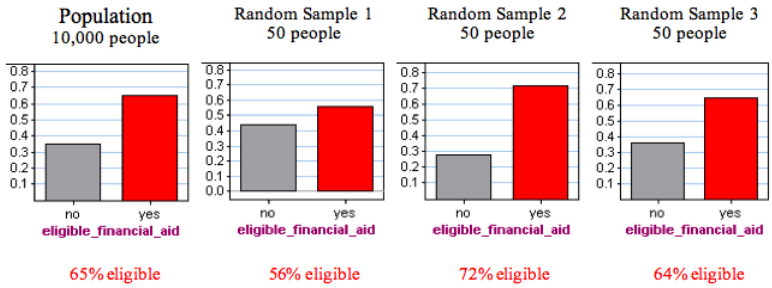
Example

For Random Samples, Size Matters

Let's compare the accuracy of random samples of different sizes.

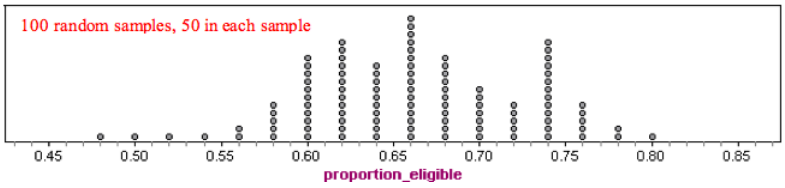
Suppose there are 10,000 students at your college. Also suppose that 65% of these students are eligible for financial aid. How accurate are random samples at predicting this population value?

To answer this question, we randomly select 50 students and determine the proportion who are eligible for financial aid. We repeat this several times. Here are the results for three random samples:



Notice that each random sample has a different result. Some results are larger than the true population value of 65%; some results are smaller than the true population value. Because there is no bias in random samples, we expect results above and below the true value to occur with similar frequency.

Now we use a simulation to take many more random samples. Again, each sample is composed of 50 randomly selected people. Here is a dotplot of the proportion who are eligible for financial aid in 100 samples. Each dot is a random sample.

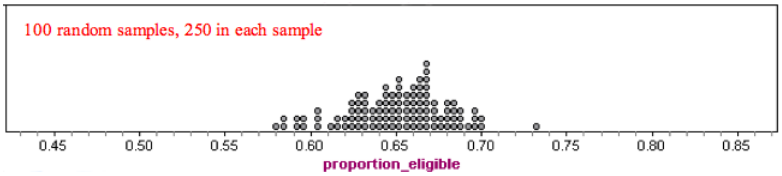


We see that the results from random samples vary from 0.48 to 0.80. Typical values range from about 0.58 to 0.74.

Note: Many samples have results below the true population value of 0.65, and many have results above 0.65. This shows that random samples are not biased. For the question *Are you eligible for financial aid?*, there is no systematic favoring of one response over another. The samples are representative of the population.

What happens when we increase the number of people in the random sample?

We increased the number of people in each sample to 250. Here is dotplot of the results from 100 of these larger random samples.



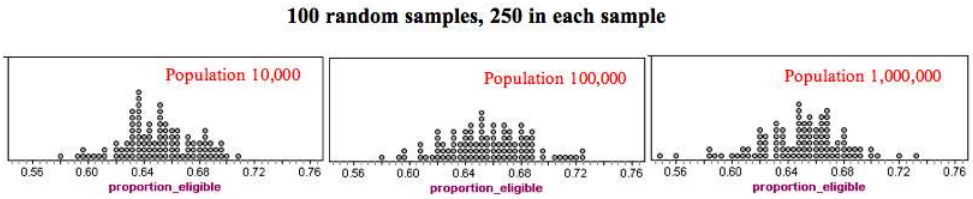
Notice there is less variability in these larger samples. Results range from about 0.58 to 0.73. Typical values range from about 0.62 to 0.68. These samples give results that are closer to the true population value of 0.65.

So what's the point? *Larger samples tend to be more accurate than smaller samples if the samples are chosen randomly.*

Comment

The precision of the sample results depends on the size of the sample, not the size of the population. The following dotplots illustrate this point. Here we selected samples with 250 people in each sample, but we *varied the size of the population*. Each dotplot contains 100 samples.

Notice that the sample results look very similar. For each population, the sample results fall between about 0.58 and 0.73. In each graph, it is common for sample results to fall between about 0.62 and 0.68.



What's the main point? *The size of the population does not affect the accuracy of a random sample as long as the population is large.*

Try It

https://assessments.lumenlearning.co...sessments/3408

https://assessments.lumenlearning.co...sessments/3828

Comment

If an attempt is made to include every individual from a population in a sample, then the investigation is called a **census**. Every 10 years, the U.S. Census Bureau conducts a population census. It attempts to collect information about every person living in the United States. However, the population census misses between 1% and 3% of the U.S. population and accidentally counts some people more than once. A full census is possible only for small populations.

Let's Summarize

- We draw a conclusion about the population on the basis of the sample.
- To draw a valid conclusion, the sample must be representative of the population. A representative sample is a subset of the population. It also reflects the characteristics of the population.
- A sample is biased if it systematically favors a certain outcome.
- Random selection eliminates bias.
- Larger samples tend to be more accurate than smaller samples if the samples are chosen randomly.
- The size of the population does not affect the accuracy of a random sample as long as the population is large.
- If an attempt is made to include every individual from a population in a sample, then the investigation is called a *census*.

Contributors and Attributions

1.10: Methods of Sampling

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of *n* individuals is equally likely to be chosen by any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in Table 1.10.2

Table 1.10.3: Class Roster

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cuningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360; 0.99832; 0.14669; 0.51470; 0.40581; 0.73381; 0.04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cuningham, and Cuarismo.

To generate random numbers:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(. Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers. If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.

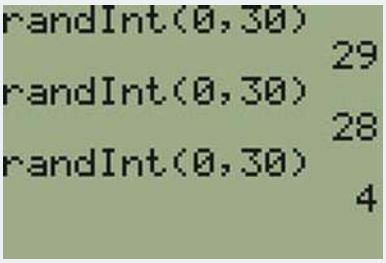


Figure 1.10.7

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Exercise 1.10.8

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Answer

a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Example 1.10.9: Calculator

You are going to use the random number generator to generate different types of samples from the data. This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6

#1	#2	#3	#4	#5	#6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick samples.

a. Create a stratified sample by column. Pick three quiz scores randomly from each column.
   - Number each row one through ten.
   - On your calculator, press Math and arrow over to PRB.
   - For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
   - Repeat for columns two through six.
   - These 18 quiz scores are a stratified sample.

b. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.
   - Press MATH and arrow over to PRB.
   - Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
   - The two numbers are for two of the columns.
   - The quiz scores (20 of them) in these 2 columns are the cluster sample.

c. Create a simple random sample of 15 quiz scores.
   - Use the numbering one through 60.
   - Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
   - Press ENTER 15 times and record the numbers.
   - Record the quiz scores that correspond to these numbers.
   - These 15 quiz scores are the systematic sample.

d. Create a systematic sample of 12 quiz scores.
   - Use the numbering one through 60.
   - Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
   - Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

Example 1.10.10

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.

f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Answer

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f.convenience

Exercise 1.10.11

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Answer

stratified

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Example 1.10.12: Sampling

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task. Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128; $87; $173; $116; $130; $204; $147; $189; $93; $153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

$50; $40; $36; $15; $50; $100; $40; $53; $22; $22

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Answer

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Answer

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

$180; $50; $150; $85; $260; $75; $180; $200; $200; $150

c. Is the sample biased?

Answer

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Exercise 1.10.12

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Answer

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The sample probably doesn't represent the entire fan base and is probably biased towards people who would prefer music.

Collaborative Exercise 1.10.8

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

a. To find the average GPA of all students in a university, use all honor students at the university as the sample.
b. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
c. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
d. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
e. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

References

1. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).
2. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).
3. Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/ga...questions.aspx (accessed May 1, 2013).
4. Data from www.bookofodds.com/Relationsh...-the-President
5. Dominic Lusinchi, "'President' Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?" Social Science History 36, no. 1: 23-54 (2012), ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).
6. "The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).
7. "Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/ga...9362004.aspx#4 (accessed May 1, 2013).
8. The Data and Story Library, lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).
9. LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/f...hts.html#focus (accessed May 1, 2013).
10. Data from San Jose Mercury News

Review

Data are individual items of information that come from a population or sample. Data may be classified as qualitative, quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

Footnotes

1. lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).
2. Scott Keeter et al., "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).
3. Frequently Asked Questions, Pew Research Center for the People & the Press, www.people-press.org/methodol...wer-your-polls (accessed May 1, 2013).

Glossary

Cluster Sampling

a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Convenience Sampling

a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Nonsampling Error

an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Random Sampling

a method of selecting a sample that gives every member of the population an equal chance of being selected.

Sampling Bias

not all members of the population are equally likely to be selected

Sampling Error

the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement

Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement

A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling

a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Stratified Sampling

a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Systematic Sampling

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

1.11: Introduction to Conducting Experiments

What you'll learn to do: Identify features of experiment design that control the effects of confounding.

In experiments, instead of assessing the values of the variables as they naturally occur, the researchers interfere and they are the ones who assign the values of the explanatory variable to the individuals. The researchers "take control" of the values of the explanatory variable because they want to see how changes in the value of the explanatory variable affect the response variable. (Note: By nature, any experiment involves at least two variables.)

The type of experiment design used, and the details of the design, are crucial, since they will determine what kind of conclusions we may draw from the results. This is especially important when we are trying to establish a cause-and-effect relationship between two variables.

Contributors and Attributions

1.12: Conducting Experiments (1 of 2)

Learning Objectives

- Identify features of experiment design that control the effects of confounding.

We now focus on experiments.

The primary goal of an **experiment** is to provide evidence for a cause-and-effect relationship between two variables. An experiment intentionally manipulates the explanatory variable in an attempt to cause an effect on the response variable. To establish a cause-and-effect relationship, we want to make sure that the explanatory variable is the only factor that impacts the response variable. We therefore attempt to get rid of all other factors that might affect the response. These other factors are called **confounding variables**.

To confound means to mix up or to confuse. Confounding variables mix up our ability to determine if the explanatory variable causes a change in the response variable. If we do not control the effects of confounding variables, the experiment does not provide evidence of a cause-and-effect relationship between the explanatory and response variables.

Researchers use two common strategies to control the effects of confounding variables:

- Direct control
- Random assignment

Example

Direct Control

Researchers compare bacteria reduction for three different hand-drying methods. In this experiment, participants handled uncooked chicken for 45 seconds, then washed their hands with one squirt of soap for 60 seconds, and then used one of three hand-drying methods. After participants completely dried their hands, researchers measured the bacteria count on their hands. The *Infectious Disease News* published the results in 2010.

In this experiment, the explanatory variable is *hand-drying method*.The response variable is *bacteria count*. Notice that the explanatory variable determines the three treatments in the experiment. Each treatment is a different hand-drying method. For this reason, the explanatory variable is also called the **treatment variable**.

In this experiment, researchers attempt to directly control the influence of three variables that could affect the bacteria count:

(1) Length of time participants handle the raw chicken.

- Direct control: All participants handle the raw chicken for 45 seconds.

(2) Amount of soap participants use.

- Direct control: All participants use one squirt of soap.

(3) Amount of time participants wash hands.

- Direct control: All participants wash their hands for 60 seconds.

Notice that the control works by stabilizing the impact of the confounding variable across the treatments. For example, the amount of soap will still influence the bacteria count. We cannot avoid this. But if all participants use the same amount of soap, then *differences* in bacteria count among the three treatments cannot be due to the amount of soap used.

Similarly, the amount of time that participants wash their hands will influence the bacteria count. But if all participants wash their hands for the same amount of time, then *differences* in bacteria count among the three treatments cannot be due to the amount of time participants washed their hands. This is what we mean when we say that the control works by stabilizing the impact of the confounding variable across the treatments.

Now we examine random assignment. Random assignment controls the effects of confounding variables that a researcher cannot control directly or that are difficult to identify in advance.

Example

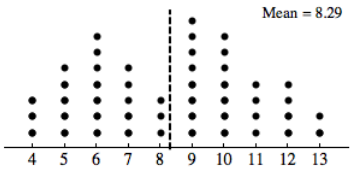
Random Assignment

Medical researchers conducted an experiment to compare two different types of surgery for children with hernias. They compared the recovery times for each type of surgery. The two surgery types are laparoscopic repair (a surgery that involves three small incisions) and open repair (a surgery that involves one large incision). Researchers identified a variety of variables that might influence recovery time, such as child's age, weight, and physical fitness level.

Let's consider one of these variables: age. How could the researchers control the impact of age on recovery time?

Direct control involves use of children of the same age. For example, researchers might use only 10-year-old children in the experiment. But it may be difficult to find enough 10-year-old children with hernias. So how do researchers create treatment groups that are similar with respect to age? One way is to assign children at random to treatment groups.

The goal of random assignment is to create similar groups with respect to age, weight, and other characteristics that might influence recovery time. To illustrate how random assignment creates similar groups, we focus on age. Here is a dotplot of the ages of the 48 children with hernias who participated in this experiment. Each dot represents a child. The average age of the 48 children is 8.29 years.



If random assignment is working, the average age for each treatment group should be about equal. We see how random assignment works in the next activity.

Click Random Assignment to randomly assign the 48 children to the two treatments. Repeat this process several times to investigate whether random assignment creates groups with similar ages. The average age is labeled as the mean and marked with a vertical line. Compare the average ages for the treatment groups.

Click here to open this simulation in its own window.

A link to an interactive elements can be found at the bottom of this page.

Try It

https://assessments.lumenlearning.co...sessments/3829

What Is the Main Point?

The goal of random assignment is to create similar treatment groups. If the groups are similar, then any differences we see in the response variable are due to the differences in treatments. In this way, random assignment controls the impact of confounding variables. Random assignment in an experiment eliminates confounding, just as random selection in a survey eliminates bias.

Comment

How do we make random assignments? We use any method that allows random chance to choose the treatment for each participant. Random assignment means that each participant has an equal chance of receiving any one of the treatment options. For example, in the hernia experiment, you could put every child's name in a hat. The first 24 names drawn get the first treatment. The rest of the children get the second treatment. You could also flip a coin. Heads means the child is assigned to the first treatment. This method could create groups with slightly different sizes, which is fine.

Try It

https://assessments.lumenlearning.co...sessments/3830

Try It

The following paragraph is from a 1999 USA Today article titled "Heart care reflects race and sex, not symptoms."

“Previous research suggested that blacks and women were less likely than whites and men to get cardiac catheterization or coronary bypass surgery for chest pain or a heart attack. Scientists blamed differences in illness severity, insurance coverage, patient preference, and health care access. The researchers eliminated those differences by videotaping actors – two black men, two white men, two black women, two white women – describing chest pain from identical scripts. They wore identical gowns, used identical hand gestures, and were taped from the same position. Researchers asked 720 primary care doctors at meetings of the American College of Physicians or the American Academy of Family Physicians to watch a tape and recommend care. The doctors thought the study focused on clinical decision making.”

Researchers rolled a four-sided die to determine which video each doctor watched.

https://assessments.lumenlearning.co...sessments/3409

https://assessments.lumenlearning.co...sessments/3410

https://assessments.lumenlearning.co...sessments/3831

https://assessments.lumenlearning.co...sessments/3832

Contributors and Attributions

1.13: Conducting Experiments (2 of 2)

Learning Objectives

- Avoid overgeneralization of experiment results.

Let's summarize what we know about experiments:

- The goal of the experiment is to provide evidence for a cause-and-effect relationship between two variables.
- A well-designed experiment controls the effects of confounding variables to isolate the effect of the explanatory variable on the response.
- Two commonly used methods for controlling the effects of confounding variables are *direct control* and *random assignment*.
- Random assignment uses random chance to assign participants to treatments. This creates similar treatment groups. With random assignment, we can be fairly confident that any differences we observe in the response of treatment groups is due to the explanatory variable. In this way, we have evidence for a cause-and-effect relationship.

Now we discuss a few more strategies that are commonly used to control the effects of confounding variables.

In an experiment, we manipulate the explanatory variable to determine if it has an effect on the response variable. Could the change we observe in the response variable happen without manipulating the explanatory variable? Maybe what we observe would have happened anyway.

For this reason, it is important to include a **control group**. A control group is a group that receives no treatment. The control group provides a baseline for comparison.

Example

Control Groups

Music and rats: In David Merrell's experiment with rats, he wanted to examine the relationship between music and the ability of rats to run a maze. He had three treatment groups: exposure to music by the heavy metal band Anthrax, exposure to music by Mozart, and no exposure to music. The group of rats that did not listen to music is the control group. Merrell's experiment lasted 1 month. With a month of practice, the rats in all the groups would probably get faster at running the maze. The control group provides a baseline for comparison. At the end of 1 month, the rats in the Mozart group were much faster at running the maze than were the rats in the other two groups. Comparison to the control group shows that the improvement in the Mozart group is not due to the rats being more experienced with the maze.

Hernia treatments for children: In this experiment, researchers compared two different surgeries. The response variable was recovery time, so it would not have made sense to have a no-treatment group. However, one type of surgery was the standard treatment for hernias, and children who received this surgery represented the control group. This group provides a baseline for comparing recovery times.

In experiments that use human participants, use of a control group may not be enough to establish whether a treatment really has an effect. A substantial amount of research shows that people respond in positive ways to treatments that have no active ingredients, a response called the **placebo effect**. A placebo is a treatment with no active ingredients, sometimes called a "sugar pill."

Example

The Placebo Effect

An article published in the *Washington Post* in 2002 illustrates the placebo effect in medical experiments.

> *After thousands of studies, hundreds of millions of prescriptions and tens of billions of dollars in sales, two things are certain about pills that treat depression: Antidepressants like Prozac, Paxil and Zoloft work. And so do sugar pills. A new analysis has found that in the majority of trials conducted by drug companies in recent decades, sugar pills have done as well as – or better than – antidepressants.…The new research may shed light on findings such as those from a trial last month that compared the herbal remedy St. John's wort against Zoloft. St. John's wort fully cured 24 percent of the depressed people who received it, and Zoloft cured 25 percent – but the placebo fully cured 32 percent.*

The placebo effect can confound the results of medical experiments. It is uncertain what is behind the placebo effect, but because people in medical experiments improve when taking a placebo, a placebo group provides a baseline for comparing treatments. We cannot eliminate the placebo effect on a treatment group. Both the placebo group and the drug group experience the placebo effect. If a treatment produces better results than a placebo, then we have evidence that the treatment (and not the placebo effect) is responsible for the improvement.

In experiments that use a placebo, participants do not know whether they are receiving the drug or a placebo. The participants are *blind* to the treatment to prevent their own beliefs about the drug (or placebo) from confounding the results.

Example

Blinding

Recall our discussion of the experiment conducted by the Women's Health Initiative to study the health implications of hormone replacement therapy. In this experiment, researchers randomly assigned over 16,000 women to one of two treatments. One group took hormones. The other group took a placebo. The experiment was also double-blind, meaning that neither the women nor the researchers knew who had which treatment.

In a *single-blind*, experiment only one of the two (either the researcher or the participants) do not know which treatment the participants receive.

Try It

https://assessments.lumenlearning.co...sessments/3833

https://assessments.lumenlearning.co...sessments/3834

To end our discussion of experiments, we consider one last question: *If an experiment is well-designed, can we generalize the results?*

Recall the hormone replacement experiment. This experiment has all of the features of a well-designed experiment:

- A large number of participants (over 16,000 women)
- Use of a placebo group
- Random assignment of women to hormone treatment or placebo
- Double-blind design

After 5 years, the group taking hormones had a higher incidence of breast cancer and heart disease. Researchers were so alarmed by the results that the experiment was ended early to prevent further harm to the health of the women participating in the hormone group.

As a result of this experiment, the use of hormone replacement therapy fell by 66%.

Yet the British Menopause Society and the International Menopause Society questioned this reaction. The Women's Health Concern, a British nonprofit group that provides independent and unbiased information about women's health, wrote:

> *It must be remembered that the WHI data on which the concerns were raised related to overweight North American women in their mid-sixties and not to the women that are treated with HRT for their menopausal symptoms in the United Kingdom, who are usually around the age of menopause, namely 45–55 years.*

The concerns expressed here do not challenge the validity of the results of the WHI experiment. Instead, they question whether the results apply to a larger group of women: women who are younger and not overweight when they go through menopause.

This is an important consideration. If our goal is to generalize the results of an experiment to a more general population, we must consider issues of sampling design as well as random assignment.

Try It

https://assessments.lumenlearning.co...sessments/3835

An Important Point about the Role of Random Chance

We now know that in an experiment we intentionally manipulate the explanatory variable to observe changes in the response variable. We use the explanatory variable to create different treatments. If we see different responses in the different treatments, we want to be able to say that the differences are the result of the explanatory variable. We must rule out other possible explanations

for the differences we observed, so we use direct control and random assignment, as well as a control group, a placebo group, or blinding as appropriate.

But none of these strategies will rule out the influence of **chance variation**. When we randomly assign participants to treatments, we produce similar groups most of the time. But there is a small chance that we will end up with treatment groups that are not similar.

For example, in the hernia experiment with children, we saw that random assignment creates two groups with average ages that are close. But there is a very small chance that we could get two groups that significantly differ in ages. This will not happen very often, but it could. And if it does happen, the results of our experiment are confounded by age.

Similarly, when we investigated how well a random sample estimates the proportion of students receiving financial aid in the population, we saw that the proportions from random samples gave good estimates – most of the time. Occasionally, a random sample did not give a good estimate. Larger random samples varied less, but they still varied.

What's the Main Point?

Good study design is important. Random selection in sampling can control bias. Random assignment in experiments can control the effects of confounding variables. But there is always a small chance, even when we randomly sample, that the results we observe in a poll do not represent the population well. Similarly, there is always a small chance, even when we use random assignment, that the differences we observe in an experiment are due to random variation and not the explanatory variable. For this reason, we have to understand how random chance behaves. This is the role of probability. We study probability later in the course, before we learn more formal statistical methods for determining if what we observe could be a result explained by chance.

Let's Summarize

- The goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.
- A well-designed experiment controls the effects of confounding variables to isolate the effect of the explanatory variable on the response.
- Two commonly used methods for controlling the effects of confounding variables are *direct control* and *random assignment*.
- Random assignment uses random chance to assign participants to treatments, which creates similar treatment groups. With random assignment, we can be fairly confident that any differences we observe in the response of treatment groups is due to the explanatory variable. In this way, we have evidence for a cause-and-effect relationship.
- Other strategies for controlling confounding variables include use of a control group, use of a placebo group, and blinding.
- A well-designed experiment provides evidence for a cause-and-effect relationship. But even in a well-designed experiment, differences in the response might be due to chance. We learn to describe chance behavior when we study probability later in the course.

Contributors and Attributions

1.14: Putting It Together- Types of Statistical Studies and Producing Data

Let's Summarize

- There are four steps in a statistical investigation:
  - Ask a question that can be answered by collecting data.
  - Decide what to measure, and then collect data.
  - Summarize and analyze.
  - Draw a conclusion, and communicate the results.
- There are two types of statistical studies:
  - Observational studies: An *observational study* observes individuals and measures variables of interest. We conduct observational studies to investigate questions about a population or about an association between two variables. An observational study alone does not provide convincing evidence of a cause-and-effect relationship.
  - Experiments: An *experiment* intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.
- In statistics, a *variable* is information we gather about individuals or objects.

Observational Studies

- In an observational study, we draw a conclusion about the population on the basis of a sample. To draw a valid conclusion, the sample must be representative of the population. A representative sample is a subset of the population. It also reflects the characteristics of the population.
- A sample is biased if it systematically favors a certain outcome. Voluntary response samples (such as Internet polls) and convenience samples (such as surveys at a mall) are biased.
- Random selection eliminates bias. In a simple random sample, everyone in the population has an equal chance of being chosen. In this way, random selection helps ensure that the sample is representative of the population.
- Larger samples tend to be more accurate than smaller samples if the samples are chosen randomly. The size of the population does not affect the accuracy of a random sample as long as the population is large.
- If an attempt is made to include every individual from a population in a sample, then the investigation is called a *census.*

Experiments

The goal of the experiment is to provide evidence for a cause-and-effect relationship between two variables. When we investigate a relationship between two variables, we identify an explanatory variable and a response variable. To establish a cause-and-effect relationship, we want to make sure the explanatory variable is the only factor that impacts the response variable. But other factors, called *confounding variables*, may also influence the response.

- A well-designed experiment takes steps to eliminate the effects of confounding variables. These steps include direct control, random assignment of people to treatment groups, use of a control or placebo, and blind conditions. Incorporating such precautions, a well-designed experiment provides convincing evidence of cause-and-effect.
- Random assignment uses random chance to assign participants to treatments, which creates similar treatment groups. With random assignment, we can be fairly confident that any differences we observe in the response of treatment groups is due to the explanatory variable. In this way, we have evidence for a cause-and-effect relationship.
- A well-designed experiment provides evidence for a cause-and-effect relationship. But even in a well-designed experiment, differences in the response might be due to chance. We learn to describe chance behavior when we study probability later in the course.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

CHAPTER OVERVIEW

2: Descriptive Statistics

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "Descriptive Statistics." You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

2.1: Organizing and Graphing Qualitative Data
2.2: Organizing and Graphing Quantitative Data
2.3: Stem-and-Leaf Displays
2.4: Measures of Central Tendency- Mean, Median and Mode
2.5: Measures of Position- Percentiles and Quartiles
2.6: Box Plots
2.7: Measures of Spread- Variance and Standard Deviation
2.8: Skewness and the Mean, Median, and Mode

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

2.1: Organizing and Graphing Qualitative Data

Learning Objectives

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

Figure 2.1.1: When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

Table 2.1.1: Fall Term 2007 (Census day)

De Anza College				Foothill College		
	Number	Percent			Number	Percent
Full-time	9,200	40.9%		Full-time	4,059	28.6%
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

- In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.
- In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.
- A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at Figures 2.1.3 and 2.1.4 and determine which graph (pie or bar) you think displays the comparisons better.

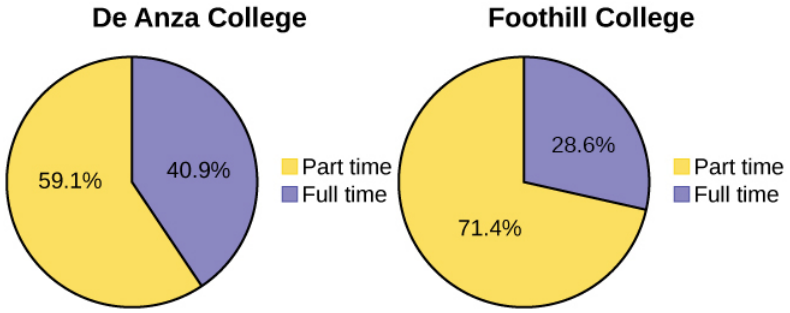


Figure 2.1.3: Pie Charts

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.

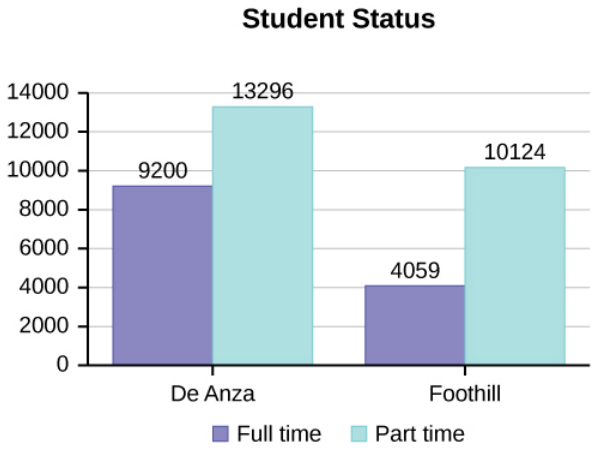


Figure 2.1.4: Bar chart

Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Table 2.1.2: De Anza College Spring 2010

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

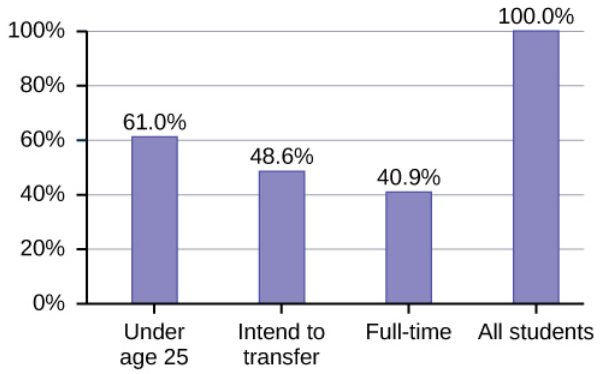


Figure 2.1.2: Bar chart of data in Table 2.1.2.

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

Table 2.1.2: Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

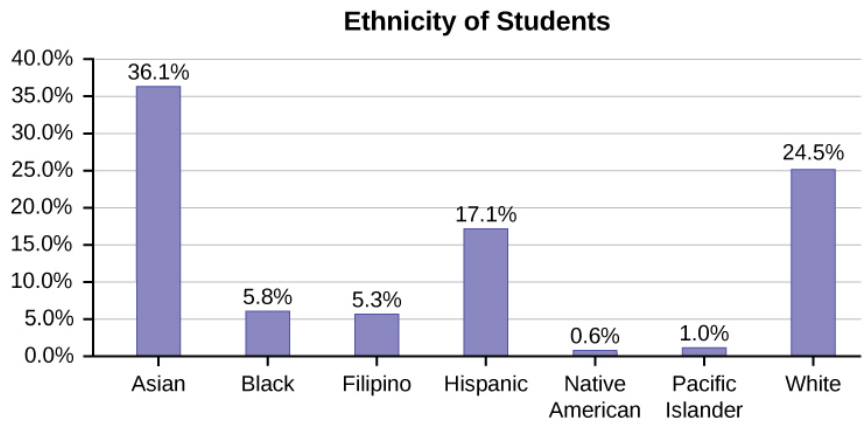


Figure 2.1.3: Enrollment of De Anza College (Spring 2010)

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been included. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in Figure 2.1.4 can be difficult to understand visually. The graph in Figure 2.1.5 is a *Pareto chart*. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

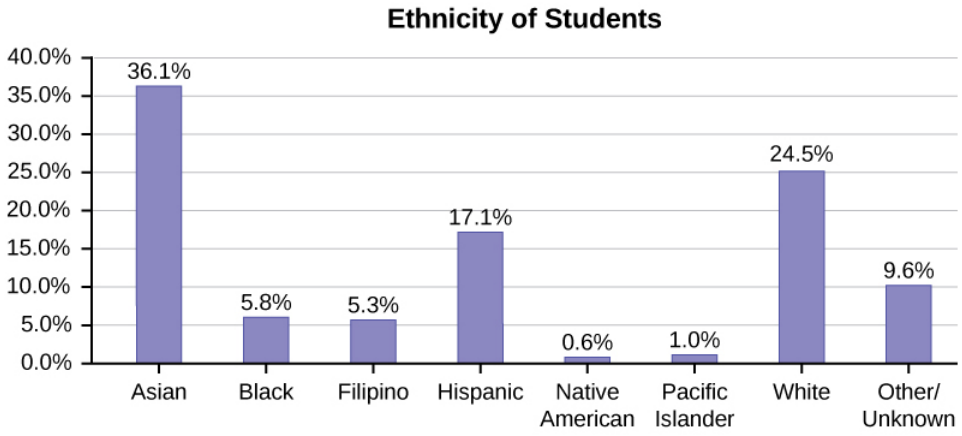


Figure 2.1.4: Bar Graph with Other/Unknown Category

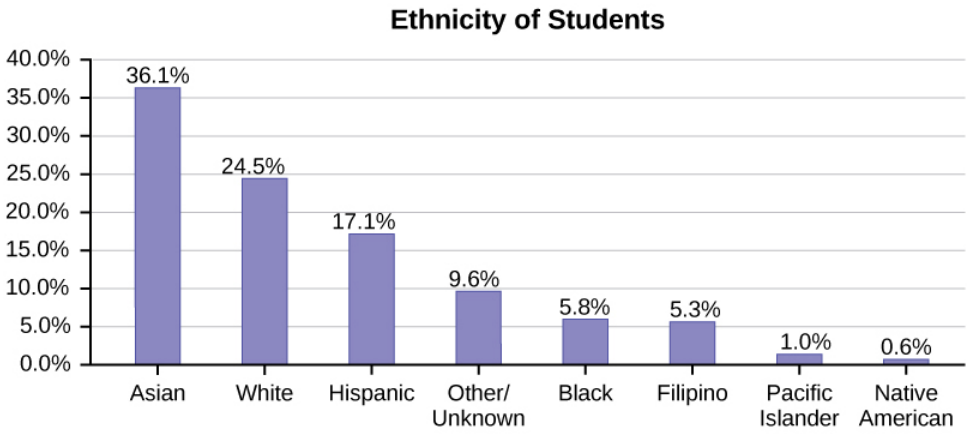


Figure 2.1.5: Pareto Chart With Bars Sorted by Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in Figure 2.1.6 is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure 2.1.6.

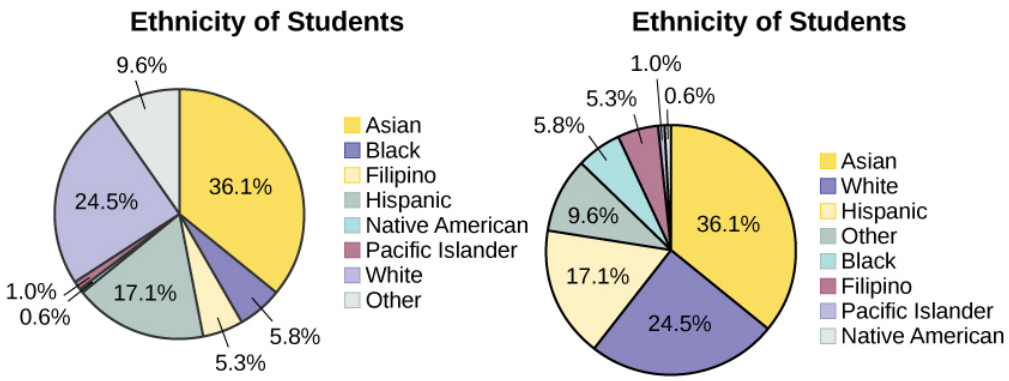


Figure 2.1.6.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

2.2: Organizing and Graphing Quantitative Data

For most of the work you do in this course, you will be working with quantitative data, and you will use a frequency table and frequency histogram to organize and graph the data. An advantage of a frequency table and frequency histogram is that they can be used to organize and display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:

5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Table lists the different data values in ascending order and their frequencies.

Table 2.2.1: Frequency Table of Student Work Hours

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

Definition: Relative Frequency

A frequency is the number of times a value of the data occurs. According to Table 2.2.1, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

Definition: Relative frequencies

A *relative frequency* is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

Table 2.2.2: Frequency Table of Student Work Hours with Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

The sum of the values in the relative frequency column of Table 2.2.2 is $\frac{20}{20}$, or 1.

Definition: Cumulative Relative Frequency

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in Table 2.2.3.

Table 2.2.3: Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	320320 or 0.15	0.15
3	5	520520 or 0.25	0.15 + 0.25 = 0.40
4	3	320320 or 0.15	0.40 + 0.15 = 0.55
5	6	620620 or 0.30	0.55 + 0.30 = 0.85
6	2	220220 or 0.10	0.85 + 0.10 = 0.95
7	1	120120 or 0.05	0.95 + 0.05 = 1.00

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Table 2.2.4 represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Table 2.2.4: Frequency Table of Soccer Player Height

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95–63.95	3	$\frac{3}{100} = 0.03$	$0.05 + 0.03 = 0.08$
63.95–65.95	15	$\frac{15}{100} = 0.15$	$0.08 + 0.15 = 0.23$
65.95–67.95	40	$\frac{40}{100} = 0.40$	$0.23 + 0.40 = 0.63$
67.95–69.95	17	$\frac{17}{100} = 0.17$	$0.63 + 0.17 = 0.80$
69.95–71.95	12	$\frac{12}{100} = 0.12$	$0.80 + 0.12 = 0.92$
71.95–73.95	7	$\frac{7}{100} = 0.07$	$0.92 + 0.07 = 0.99$
73.95–75.95	1	$\frac{1}{100} = 0.01$	$0.99 + 0.01 = 1.00$
	Total = 100	**Total = 1.00**	

The data in this table have been **grouped** into the following intervals:

- 61.95 to 63.95 inches
- 63.95 to 65.95 inches
- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches
- 73.95 to 75.95 inches

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players

whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Collaborative Exercise 2.2.7

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

a. What percentage of the students in your class have no siblings?
b. What percentage of the students have from one to three siblings?
c. What percentage of the students have fewer than three siblings?

Example 2.2.7

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. Table 2.2.6 was produced:

Table 2.2.6: Frequency of Commuting Distances

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{3}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

a. Is the table correct? If it is not correct, what is wrong?
b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
c. What fraction of the people surveyed commute five or seven miles?
d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Answer

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.
c. $\frac{5}{19}$
d. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample.(Remember, frequency is defined as the number of times an answer occurs.) If:

- f is frequency
- n is total number of data values (or the sum of the individual frequencies), and
- RF is relative frequency,

then:

$$RF = \frac{f}{n} \tag{2.2.1}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, f = 3, n = 40, and RF = fn = 340 = 0.075. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram, first decide how many bars or intervals, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A convenient starting point is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is $6.05 (6.1 – 0.05 = 6.05)$. We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is $1.495 (1.5 – 0.005 = 1.495)$. If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is $0.9995 (1.0 – 0.0005 = 0.9995)$. If all the data happen to be integers and the smallest value is two, then a convenient starting point is $1.5 (2 – 0.5 = 1.5)$. Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

Example 2.2.1

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5

66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5

68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5

70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71

72; 72; 72; 72.5; 72.5; 73; 73.5

74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

60 – 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74, so 74 + 0.05 = 74.05 is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

$$\frac{74.05 - 59.95}{8} = 1.76 \quad (2.2.2)$$

We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the width of a bar or class interval is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.

The following histogram displays the heights on the *x*-axis and relative frequency on the *y*-axis.

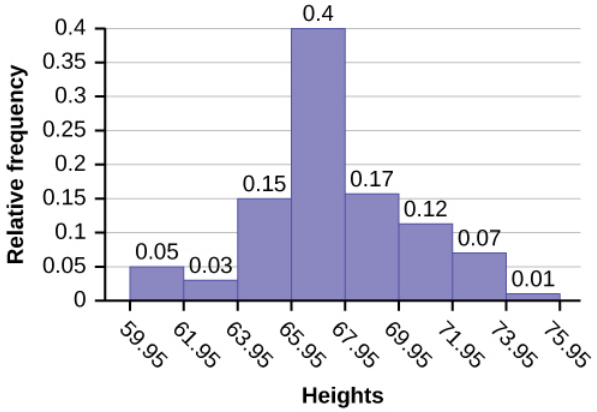


Figure 2.2.1: Histogram of something

Example 2.2.2

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____ .

Answer

Calculate the number of bars as follows:

$$\frac{6.5-0.5}{\text{number of bars}} = 1$$

where 1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the *x*-axis and the frequency on the *y*-axis.

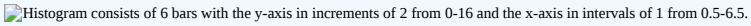


Figure 2.2.2.

Example 2.2.3

Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends

9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Answer

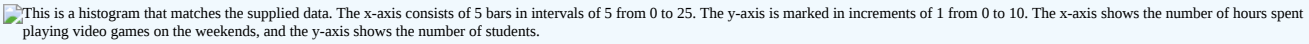


Figure 2.2.3.

Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons. To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the *x*-axis and *y*-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Example 2.2.4

A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

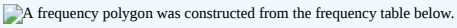


Figure 2.2.4.

The first label on the *x*-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 54.5 represents the next interval, or the first "real" interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don't have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

Constructing a Time Series Graph

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

Example 2.2.6

The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Answer

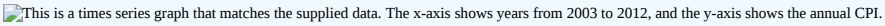


Figure 2.2.7.

Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on *y*-axis with the frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.Glossary

WeBWorK Problems

References

1. Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker
2. "Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at www.scholastic.com/teachers/a...-us-presidents (accessed April 3, 2013).
3. "Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).

4. “Food Security Statistics.” Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).
5. “Consumer Price Index.” United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).
6. “CO2 emissions (kt).” The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).
7. “Births Time Series Data.” General Register Office For Scotland, 2013. Available online at www.gro-scotland.gov.uk/stati...me-series.html (accessed April 3, 2013).
8. “Demographics: Children under the age of 5 years underweight.” Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).
9. Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.
10. “Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Frequency

the number of times a value of the data occurs

Histogram

a graphical representation in $x - y$ form of the distribution of data in a data set; x represents the data and y represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

2.3: Stem-and-Leaf Displays

One simple graph, the *stem-and-leaf graph* or *stemplot*, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a final significant digit. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 2.3.1

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem-and-Leaf Graph

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61

Construct a stem plot for the data.

Answer

Stem	Leaf
3	2 2 3 4 8
4	0 2 2 3 4 6 7 7 8 8 8 9
5	0 0 1 2 2 2 3 4 6 7 7
6	0 1

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in Example, the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are

connected using line segments.

Example 2.3.7

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in Table and in Figure.

Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

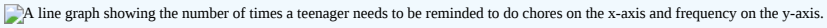


Figure 2.3.1

Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in Example 2.3.9 has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

Example 2.3.9

By the end of 2011, Facebook had over 146 million users in the United States. Table shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Answer

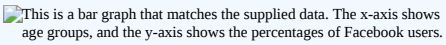


Figure 2.3.3.

Exercise 2.3.10

The population in Park City is made up of children, working-age adults, and retirees. Table shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Answer

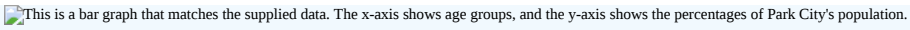


Figure 2.3.4.

Summary

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

WeBWorK Problems

References

1. Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics*, 2011. Available online at www.kenburbary.com/2011/03/fa...-statistics-2/ (accessed August 21, 2013).
2. “9th Annual AP Report to the Nation.” CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goa...omoting-equity (accessed September 13, 2013).
3. “Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.
- CUNY OER WeBWorK Fellows

2.4: Measures of Central Tendency- Mean, Median and Mode

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the median. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

> The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced “x bar”): $\bar{x}$.

The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

$$1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4$$

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \quad (2.4.1)$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7 \quad (2.4.2)$$

In the second calculation, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression

$$\frac{n+1}{2} \quad (2.4.3)$$

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then

$$\frac{n+1}{2} = \frac{97+1}{2} = 49. \quad (2.4.4)$$

The median is the 49th value in the ordered data. If the total number of data values is 100, then

$$\frac{n+1}{2} = \frac{100+1}{2} = 50.5. \quad (2.4.5)$$

The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 2.4.1

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47

Calculate the mean and the median.

Answer

The calculation for the mean is:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+\ldots+35+37+40+(44)(2)+47]}{40} = 23.6 \quad (2.4.6)$$

To find the median, M, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5 \tag{2.4.7}$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47

$$M = \frac{24+24}{2} = 24 \tag{2.4.8}$$

Example 2.4.2

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean or the median?

Solution

$$\bar{x} = \frac{5,000,000+49(30,000)}{50} = 129,400 \tag{2.4.9}$$

$M = 30,000$

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example 2.4.3

Statistics exam scores for 20 students are as follows:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Find the mode.

Answer

The most frequent score is 72, which occurs five times. Mode = 72.= 7.

Example 2.4.4

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

> The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample is very likely to get closer and closer to μ. This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a sampling distribution as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A statistic is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\bar{x}$ is an example of a statistic which estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean:

$$mean = \frac{\text{data sum}}{\text{number of data values}}. \tag{2.4.10}$$

We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is

$$\frac{\text{lower boundary}+\text{upper boundary}}{2}. \tag{2.4.11}$$

We can now modify the mean definition to be

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{2.4.12}$$

where f is the frequency of the interval and m is the midpoint of the interval.

Example 2.4.5

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3

Grade Interval	Number of Students
86.5–92.5	4
92.5–98.5	1

Solution

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

- Calculate the sum of the product of each interval frequency and midpoint.
  $\sum fm53.25(1)+59.5(0)+65.5(4)+71.5(4)+77.5(2)+83.5(3)+89.5(4)+95.5(1)=1460.25$
- $\mu=\dfrac{\sum fm}{\sum f}=\dfrac{1460.25}{19}=76.86$

WeBWorK Problems

References

1. Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).
2. “Demographics: Obesity – adult prevalence rate.” Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

Review

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occuring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

Formula Review

$$\mu=\frac{\sum fm}{\sum f} \tag{2.4.13}$$

where f = interval frequencies and m = interval midpoints.

Glossary

Frequency Table

a data representation in which grouped data is displayed along with the corresponding frequencies

Mean

a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median

a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint

the mean of an interval in a frequency table

Mode

the value that appears most frequently in a set of data

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

2.5: Measures of Position- Percentiles and Quartiles

The common measures of location are **quartiles** and **percentiles.** Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The median, M, is called both the second quartile and the 50th percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1

Ordered from smallest to largest:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8+7.2}{2}=7 \tag{2.5.1}$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.

1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, $Q3$, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 - Q_1 \tag{2.4.1}$$

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5)(IQR) below the first quartile or more than (1.5)(IQR) above the third quartile**. Potential outliers always require further investigation.

Definition: Outliers

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Example 2.4.1

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Answer

Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$$M = 488,800$$

$$Q_1 = \frac{230,500 + 387,000}{2} = 308,750$$

$$Q_3 = \frac{639,000 + 659,000}{2} = 649,000$$

$$IQR = 649,000 - 308,750 = 340,250$$

$$(1.5)(IQR) = (1.5)(340,250) = 510,375$$

$$Q_1 - (1.5)(IQR) = 308,750 - 510,375 = -201,625$$

$$Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375$$

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Example 2.4.2

For the two data sets in the test scores example, find the following:

a. The interquartile range. Compare the two interquartile ranges.
b. Any outliers in either set.

Answer

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

a. The *IQR* for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$

The *IQR* for the night group is $Q_3 - Q_1 = 89 - 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.

b. Day class outliers are found using the IQR times 1.5 rule. So,

- $Q_1 - IQR(1.5) = 56 - 26.5(1.5) = 16.25$
- $Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$

Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

- $Q_1 - IQR(1.5) = 78 - 11(1.5) = 61.5$
- $Q_3 + IQR(1.5) = 89 + 11(1.5) = 105.5$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Example 2.4.3

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Example 2.4.4

Using Table:

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

Solution

Using the data from the frequency table, we have:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
b. The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.
c. Q_1 is also the 25th percentile. The 25th percentile location calculation: $P_{25} = 0.25(50) = 12.5 \approx 13$ the 13th data value. Thus, the 25th percentile is six.

COLLABORATIVE STATISTICS

Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions:

a. How many students were surveyed?
b. What kind of sampling did you do?
c. Construct two different histograms. For each, starting value = _____ ending value = ____.
d. Find the median, first quartile, and third quartile.
e. Construct a table of the data to find the following:
   i. the 10th percentile
   ii. the 70th percentile
   iii. the percent of students who own less than four sweaters

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the kth percentile. Here is one of them.

- $k =$ the kth percentile. It may or may not be part of the data.
- $i =$ the index (ranking or position of a data value)
- $n =$ the total number of data

Order the data from smallest to largest.

Calculate $i = \frac{k}{100}(n+1)$ i=k100(n+1)

If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.

If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example 2.4.5

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70th percentile.
b. Find the 83rd percentile.

Solution

a.
   - $k = 70$
   - i = the index
   - $n = 29$

   $i = \frac{k}{100}(n+1) = \frac{70}{100}(29+1) = 21$. Twenty-one is an integer, and the data value in the 21st position in the ordered data set is 64. The 70th percentile is 64 years.

b.
   - k = 83rd percentile
   - $i = theindex$
   - $n = 29$

   $i = \frac{k}{100}(n+1) = (\frac{83}{100})(29+1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24th position is 71 and the age in the 25th position is 72. Average 71 and 72. The 83rd percentile is 71.5 years.

Note 2.4.2

You can calculate percentiles using calculators and computers. There are a variety of online calculators.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- $x =$ the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- $y =$ the number of data values equal to the data value for which you want to find the percentile.
- $n =$ the total number of data.
- Calculate $\frac{x+0.5y}{n}(100)$. Then round to the nearest integer.

Example 2.4.6

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile for 58.
b. Find the percentile for 25.

Solution

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

$x = 18$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{18+0.5(1)}{29}(100) = 63.80$. 58 is the 64th percentile.

b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

$x = 3$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{3+0.5(1)}{29}(100) = 12.07$. Twenty-five is the 12thpercentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

GUIDELINE

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Answer

- Twenty-five percent of students finished the exam in 35 minutes or less.

- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Answer

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Example 2.4.9

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is seven units. Interpret the 30th percentile in the context of this situation.

Answer

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Example 2.4.10

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

- Min = 0
- Q_1 = 20
- Med = 40
- Q_3 = 60
- Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120 \tag{2.5.2}$$

.

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- Min = 0
- Q_1 = 20
- Q_3 = 60

- Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

WeBWorK Problems

References

1. Cauchon, Dennis, Paul Overberg. “Census data shows minorities now a majority of U.S. births.” USA Today, 2012. Available online at usatoday30.usatoday.com/news/...sus/55029100/1 (accessed April 3, 2013).
2. Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).
3. “1990 Census.” United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).
4. Data from *San Jose Mercury News*.
5. Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Review

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other obeservations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile,the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the the 75th percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

Formula Review

$$i = \frac{k}{100}(n+1)$$

where i = the ranking or position of a data value,

- k = the kth percentile,
- n = total number of data.

Expression for finding the percentile of a data value: $\left(\frac{x+0.5y}{n}\right)(100)$

where $x =$ the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

$y =$ the number of data values equal to the data value for which you want to find the percentile,

$n =$ total number of data

Glossary

Interquartile Range

or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Outlier

an observation that does not fit the rest of the data

Percentile

a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50^{th} percentile. The first and third quartiles are the 25^{th} and the 75^{th} percentiles, respectively.

Quartiles

the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

2.6: Box Plots

Box plots (also called *box-and-whisker plots* or *box-whisker plots*) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately *the middle 50 percent of the data fall inside the box*. The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

> You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider, again, this dataset.

1; 1; 2; 2; 4; 6; 6;.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

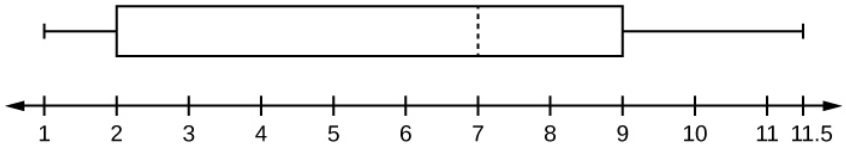


Figure 2.6.1

The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

> It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

Example 2.6.1

The following data are the heights of 40 students in a statistics class.

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77

Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- $Q1$: First quartile = 64.5
- $Q2$: Second quartile or median= 66
- $Q3$: Third quartile = 70

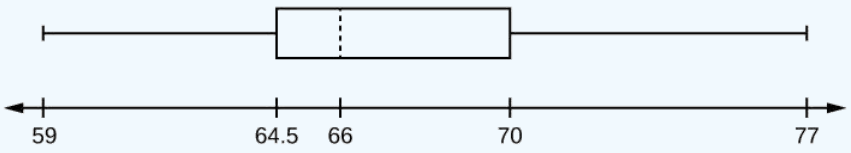


Figure 2.6.2

a. Each quarter has approximately 25% of the data.

b. The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
c. $\text{Range} = \text{maximum value} - \text{the minimum value} = 77 - 59 = 18$
d. Interquartile Range: $IQR = Q_3 - Q_1 = 70 - 64.5 = 5.5$.
e. The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
f. The middle 50% (middle half) of the data has a range of 5.5 inches.

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:

Figure 2.6.4

In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and five, inclusive. At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Example 2.6.2

Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

a. Find the smallest and largest values, the median, and the first and third quartile for the day class.
b. Find the smallest and largest values, the median, and the first and third quartile for the night class.
c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.
e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Answer

a.
- Min = 32
- Q_1 = 56
- M = 74.5
- Q_3 = 82.5
- Max = 99

b.
- Min = 25.5
- Q_1 = 78
- M = 81
- Q_3 = 89
- Max = 98

c. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class:

d.

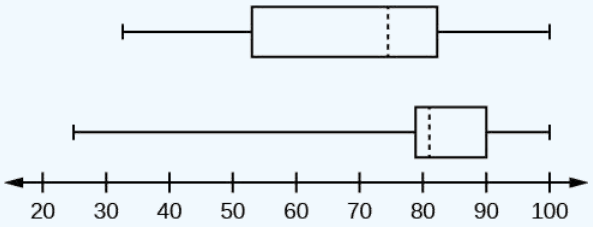


***Figure* 2.6.5**

e. The first data set has the wider spread for the middle 50% of the data. The *IQR*for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

Example 2.6.3

Graph a box-and-whisker plot for the data values shown.

10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q_1: 15
- Med: 95
- Q_3: 490
- Max: 790

The following graph shows the box-and-whisker plot.

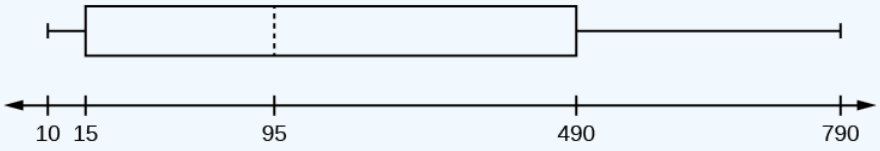


Figure 2.6.7

References

1. Data from *West Magazine*.

Review

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

WeBWorK Problems

Glossary

Box plot

a graph that gives a quick picture of the middle 50% of the data

First Quartile

the value that is the median of the of the lower half of the ordered data set

Frequency Polygon

looks like a line graph but uses intervals to display ranges of large amounts of data

Interval

also called a class interval; an interval represents a range of data and is used when displaying large data sets

Paired Data Set

two data sets that have a one to one relationship so that:

- both data sets are the same size, and
- each data point in one data set is matched with exactly one point from the other set.

Skewed

used to describe data that is not symmetrical; when the right side of a graph looks “chopped off” compared the left side, we say it is “skewed to the left.” When the left side of the graph looks “chopped off” compared to the right side, we say the data is “skewed to the right.” Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

2.7: Measures of Spread- Variance and Standard Deviation

[NOTE from VS: The following is pulled from Shafer and Zhang]

Look at the two data sets in Table 2.7.1 and the graphical representation of each, called a *dot plot*, in Figure 2.7.1.

Table 2.7.1*: Two Data Sets*

Data Set I:	40	38	42	40	39	39	43	40	39	40
Data Set II:	46	37	40	33	42	36	40	47	34	45

The two sets of ten measurements each center at the same value: they both have mean, median, and mode equal to 40. Nevertheless a glance at the figure shows that they are markedly different. In Data Set I the measurements vary only slightly from the center, while for Data Set II the measurements vary greatly. Just as we have attached numbers to a data set to locate its center, we now wish to associate to each data set numbers that measure quantitatively how the data either scatter away from the center or cluster close to it. These new quantities are called measures of variability, and we will discuss three of them.

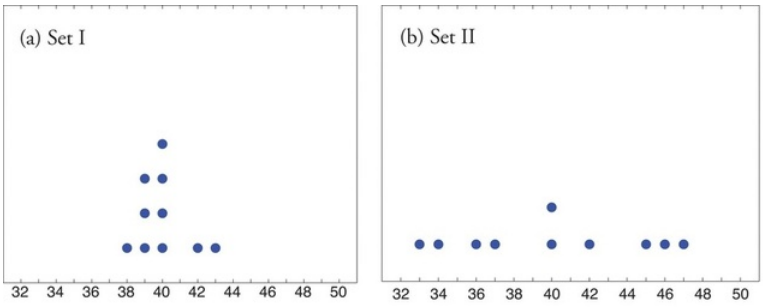


Figure 2.7.1: Dot Plots of Data Sets

The Range

First we discuss the simplest measure of variability.

Definition: range

The *range* R of a data set is difference between its largest and smallest values

$$R = x_{\max} - x_{\min} \tag{2.7.1}$$

where $x_{\max}$ is the largest measurement in the data set and $x_{\min}$ is the smallest.

Example 2.7.1: Identifyig the Range of a dataset

Find the range of each data set in Table $2.7.1$.

Solution:

- For Data Set I the maximum is 43 and the minimum is 38, so the range is $R = 43 - 38 = 5$.
- For Data Set II the maximum is 47 and the minimum is 33, so the range is $R = 47 - 33 = 14$.

The range is a measure of variability because it indicates the size of the interval over which the data points are distributed. A smaller range indicates less variability (less dispersion) among the data, whereas a larger range indicates the opposite. The range is very limited in the information it gives us, as it is only based on the largest and smallest values. Anything can happen in between and the range tells us nothing about these. In order to get information about how all of the data points are spread out we can compare each one to the mean. We do this with the "Variance" and "Standard Deviation".

The Variance and the Standard Deviation

The other two measures of variability that we will consider are the Variance and the Standard Deviation. They are intimately connected, as the standard deviation is just the square root of the variance. The word "deviation" gives us the clue of what we are trying to do. In order to measure how much variation there is in the data, we use the mean as the central value and then calculate all of the differences ("deviations") of each data value from the mean. The Variance is easier to calculate because it does not involve the square root. It has a drawback in that the quantities used are squared, so it will not represent the correct units for the data. The Standard Deviation takes the square root of the Variance and so the squared units are returned to regular units (such as inches, pounds and so forth based on the sampled data0.

Calculating the Standard Deviation

If x is a number, then the difference "x – mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$.

To calculate the standard deviation, we need to calculate the variance first, and then take the square root. The variance is the **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by ***n* – 1**, one less than the number of items in the sample.

In summary, the procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are the same except that for the sample we divide by "sample size - 1: n-1" and for the population we divide by "Population size N". Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \tag{2.7.2}$$

For the sample standard deviation, the denominator is $n - 1$, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}} \tag{2.7.3}$$

For the population standard deviation, the denominator is N, the number of items in the population.

The Sample Variance is the calculation before taking the square root.

Definition: sample variance and sample Standard Deviation

The *sample variance* of a set of n sample data is the number $\mathbf{s^2}$ defined by the formula

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} \tag{2.7.4}$$

An algebraically equivalent formula is sometimes used, because the calculations are easier to perform:

$$s^2 = \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1} \tag{2.7.5}$$

The square root **s** of the sample variance is called the *sample standard deviation* of a set of n *sample data* . It is given by the formulas

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1}}. \tag{2.7.6}$$

Although the first formula in each case looks less complicated than the second, the latter is easier to use in hand computations, and is called a *shortcut formula*.

Example 2.7.2: Identifying the Variance and Standard Deviation of a Dataset

Find the sample variance and the sample standard deviation of Data Set II in Table 2.7.1

Solution

To use the defining formula (the first formula) in the definition we first compute for each observation x its deviation $x - \bar{x}$ from the sample mean. Since the mean of the data is $\bar{x} = 40$, we obtain the ten numbers displayed in the second line of the supplied table

x	46	37	40	33	42	36	40	47	34	45
$x - \bar{x}$	-6	-3	0	-7	2	-4	0	7	-6	5

Thus

$$\sum (x-\bar{x})^2 = 6^2 + (-3)^2 + 0^2 + (-7)^2 + 2^2 + (-4)^2 + 0^2 + 7^2 + (-6)^2 + 5^2 = 224$$

so the variance is

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{224}{9} = 24.\bar{8}$$

and the standard deviation is

$$s = \sqrt{24.\bar{8}} \approx 4.99$$

The student is encouraged to compute the ten deviations for Data Set I and verify that their squares add up to 20, so that the sample variance and standard deviation of Data Set I are the much smaller numbers

$$s^2 = 20/9 = 2.\bar{2} \tag{2.7.7}$$

and

$$s = 20/9 \approx 1.49 \tag{2.7.8}$$

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

WeBWorK Problems

The number of standard deviations a data value is from the mean can be used as a measure of the closeness of a data value to the mean. Because the standard deviation measures the spread of the data this gives a uniform measure for any data set.

For example: suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes.

Rosa waits for seven minutes:

- This is two minutes longer than the average wait time.
- Two minutes is the same as one standard deviation.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- This is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A "rule of thumb" is that more than two standard deviations away from the average is considered "far from the average". In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because $5+(1)(2)=7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because $5+(-2)(2)=1$.

Figure 2.7.1

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: $1=5+(-2)(2)$.

The equation value = mean + (#ofSTDEVs)(standard deviation) can be expressed for a sample and for a population.

- sample:

$$x=\bar{x}+(\#\text{ofSTDEV})(s) \tag{2.7.9}$$

- Population:

$$x=\mu+(\#\text{ofSTDEV})(s) \tag{2.7.10}$$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol $\bar{x}$ is the sample mean and the Greek symbol μ is the population mean.

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example 2.7.1

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of *n* = 20 fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating *s*.

Data	Freq.	Deviations	*Deviations*^{2}	(Freq.)(*Deviations*^{2})
x	*f*	$(x - \bar{x})$	$(x - \bar{x})2$	$(f)(x - \bar{x})2$
9	1	9 – 10.525 = –1.525	(–1.525)2 = 2.325625	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(–1.025)^2$ = 1.050625	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(–0.525)^2$ = 0.275625	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(–0.025)^2$ = 0.000625	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 × 0.950625 = 2.851875
				The total is 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$$s^2 = \frac{9.7375}{20 - 1} = 0.5125$$

The **sample standard deviation** *s* is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = 0.715891$$

and this is rounded to two decimal places, $s = 0.72$.

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation or a calculator or computer.
- For a sample: $x = \bar{x}$ + (#ofSTDEVs)(*s*)
- For a population: $x = \mu$ + (#ofSTDEVs)σ
- For this example, use $x = \bar{x}$ + (#ofSTDEVs)(*s*) because the data is from a sample

a. Verify the mean and standard deviation on your calculator or computer.
b. Find the value that is one standard deviation above the mean. Find ($\bar{x}$ + 1s).
c. Find the value that is two standard deviations below the mean. Find ($\bar{x}$ – 2s).
d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

a.
- Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.

- $\bar{x}$ = 10.525
- Use Sx because this is sample data (not a population): Sx=0.715891

b. $(\bar{x}+1s) = 10.53 + (1)(0.72) = 11.25$
c. $(\bar{x}-2s) = 10.53 - (2)(0.72) = 9.09$
d.
- $(\bar{x}-1.5s) = 10.53 - (1.5)(0.72) = 9.45$
- $(\bar{x}+1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For Example 2.7.1, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one ($n - 1$). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by ($n - 1$) gives a better estimate of the population variance.

Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Example 2.7.2

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
   i. The sample mean
   ii. The sample standard deviation
   iii. The median
   iv. The first quartile
   v. The third quartile
   vi. *IQR*
c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Answer

a. See Table

b. i. The sample mean = 73.5
ii. The sample standard deviation = 17.9
iii. The median = 73
iv. The first quartile = 61
v. The third quartile = 90
vi. *IQR* = 90 – 61 = 29

c. The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is $(100.5 - 32.5)$ divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, $32.5 + 13.6 = 46.1$, $46.1 + 13.6 = 59.7$, $59.7 + 13.6 = 73.3$, $73.3 + 13.6 = 86.9$, $86.9 + 13.6 = 100.5 =$ the ending value; No data values fall on an interval boundary.

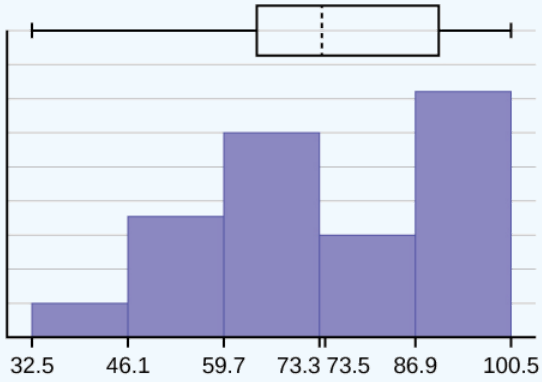


Figure 2.7.2.

The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater $(73 - 33 = 40)$ than the spread in the upper 50% $(100 - 73 = 27)$. The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (*IQR* = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{2.7.11}$$

where f interval frequencies and $m =$ interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how “unusual” individual data is compared to the mean.

Example 2.7.3

Find the standard deviation for the data in Table 2.7.3.

Table 2.7.3

Class	Frequency, f	Midpoint, m	m^2	$\bar{x}$	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

For this data set, we have the mean, $\bar{x}$ = 7.58 and the standard deviation, s_x = 3.5. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since 7.58 – 3.5 – 3.5 = 0.58. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{f(m-\bar{x})^2}{n-1}}$ where s_x = sample standard deviation, $\bar{x}$ = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#\text{ofSTDEVs} = \dfrac{\text{value-mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	$x = \bar{x} + zs$	
Population	$x = \mu + z\sigma$	

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Answer

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \#\text{ofSTDEVs} = \left(\frac{\text{value-mean}}{\text{standard deviation}}\right) = \left(\frac{x+\mu}{\sigma}\right)$$

For John,

$$z = \#\text{ofSTDEVs} = \left(\frac{2.85 - 3.0}{0.7}\right) = -0.21$$

For Ali,

$$z = \#\text{ofSTDEVs} = \left(\frac{77 - 80}{10}\right) = -0.3$$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's *z*-score of –0.21 is higher than Ali's *z*-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

1. Data from Microsoft Bookshelf.
2. King, Bill."Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x-\bar{x})^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}}$. .

Formula Review

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} \tag{2.7.12}$$

where s_x sample standard deviation and $\bar{x}$ = sample mean

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

Glossary

Standard Deviation
a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation.

Variance
mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Contributors and Attributions

2.8: Skewness and the Mean, Median, and Mode

Consider the following data set.

4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.

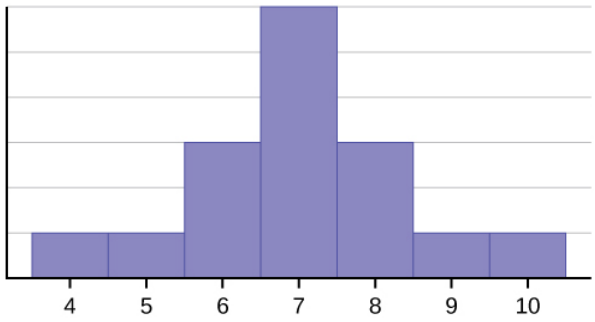


Figure 2.8.1

The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.

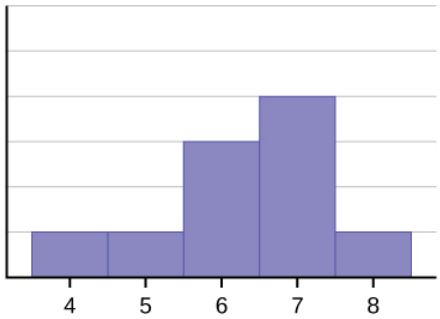


Figure 2.8.2

The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10, is also not symmetrical. It is **skewed to the right**.

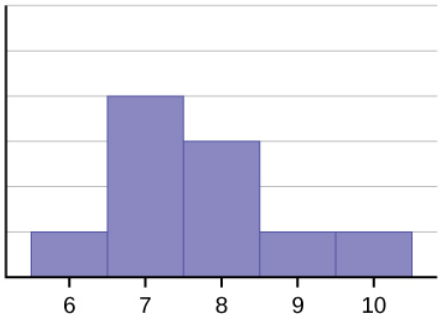


Figure 2.8.3

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

> *Generally, if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.*

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Example 2.8.1

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

- Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2
- Davis: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1
- Maris: 2; 3; 4; 4; 4; 6; 6; 6; 8; 3

a. Make a dot plot for the three authors and compare the shapes.
b. Calculate the mean for each.
c. Calculate the median for each.
d. Describe any pattern you notice between the shape and the measures of center.

Solution

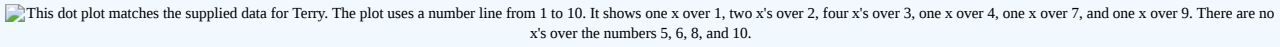
a. 

Figure 2.8.4: Terry's distribution has a right (positive) skew.

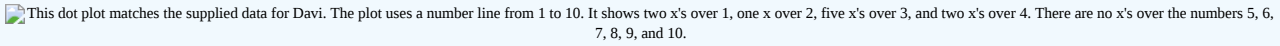


Figure 2.8.5: Davis' distribution has a left (negative) skew

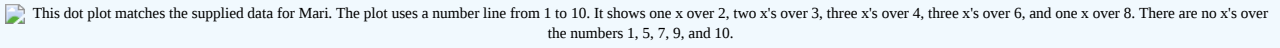


Figure 2.8.6: Maris' distribution is symmetrically shaped.

b. Terry's mean is 3.7, Davis' mean is 2.7, Maris' mean is 4.6.
c. Terry's median is three, Davis' median is three. Maris' median is four.
d. It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **left (or negative) skewed** distribution has a shape like Figure 2.8.2. A **right (or positive) skewed** distribution has a shape like Figure 2.8.3. A **symmetrical** distribution looks like Figure 2.8.1.

Contributors and Attributions

CHAPTER OVERVIEW

3: Examining Relationships- Quantitative Data

3.1: Why It Matters- Examining Relationships- Quantitative Data
3.2: Linear Regression (4 of 4)
3.3: Introduction to Assessing the Fit of a Line
3.4: Assessing the Fit of a Line (1 of 4)
3.5: Assessing the Fit of a Line (2 of 4)
3.6: Assessing the Fit of a Line (3 of 4)
3.7: Assessing the Fit of a Line (4 of 4)
3.8: Putting It Together- Examining Relationships- Quantitative Data
3.9: StatTutor- Academic Performance
3.10: Assignment- Scatterplot
3.11: Assignment- Linear Relationships
3.12: Introduction to Scatterplots
3.13: Assignment- Linear Regression
3.14: Scatterplots (1 of 5)
3.15: Scatterplots (2 of 5)
3.16: Scatterplots (3 of 5)
3.17: Scatterplots (4 of 5)
3.18: Scatterplots (5 of 5)
3.19: Introduction to Linear Relationships
3.20: Linear Relationships (1 of 4)
3.21: Linear Relationships (2 of 4)
3.22: Linear Relationships (3 of 4)
3.23: Linear Relationships (4 of 4)
3.24: Introduction to Association vs Causation
3.25: Causation and Lurking Variables (1 of 2)
3.26: Causation and Lurking Variables (2 of 2)
3.27: Introduction to Linear Regression
3.28: Linear Regression (1 of 4)
3.29: Linear Regression (2 of 4)
3.30: Linear Regression (3 of 4)

3.1: Why It Matters- Examining Relationships- Quantitative Data

Why learn how to analyze data by examining the relationships within quantitative data?

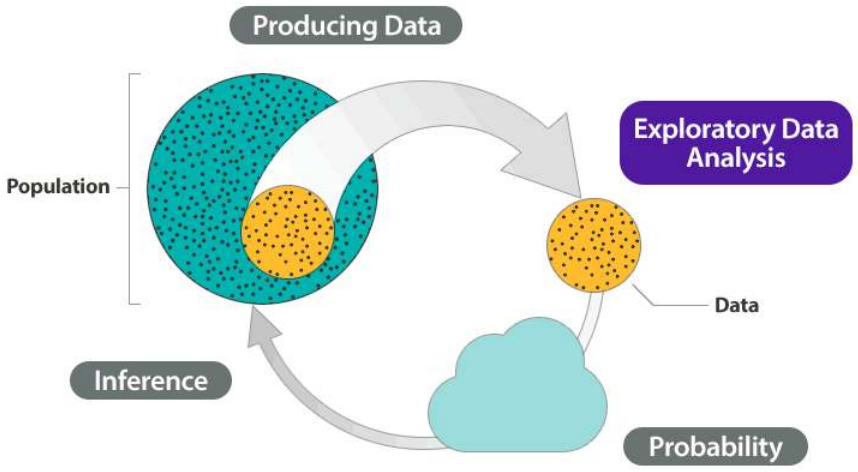
Before we begin *Examining Relationships: Quantitative Data*, let's see how the new ideas in this module relate to what we learned in the previous modules, *Types of Statistical Studies and Producing Data* and *Summarizing Data Graphically and Numerically*.

Recall the Big Picture:

We begin a statistical investigation with a research question. The investigation proceeds with the following steps:

- Produce Data: Determine what to measure, then collect the data. ← **Types of Statistical Studies and Producing Data**
- Explore the Data: Analyze and summarize the data. ← **Summarizing Data Graphically and Numerically, Examining Relationships: Quantitative Data**
- Draw a Conclusion: Use the data, probability, and statistical inference to draw a conclusion about the population.

Types of Statistical Studies and Producing Data focused on methods for collecting reliable data. *Summarizing Data Graphically and Numerically* focused on summarizing and analyzing data for a quantitative variable. In this module, we focus on summarizing and analyzing the relationship between two quantitative variables. In the Big Picture of Statistics, the material in *Examining Relationships: Quantitative Data* is still part of exploratory data analysis.



Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.2: Linear Regression (4 of 4)

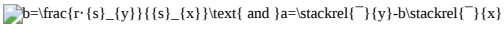
Learning Objectives

- For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

In the previous activity we used technology to find the least-squares regression line from the data values.

We can also find the equation for the least-squares regression line from summary statistics for x and y and the correlation.

If we know the mean and standard deviation for x and y, along with the correlation (r), we can calculate the slope b and the starting value a with the following formulas:



$$b=\frac{r\cdot{s}_{y}}{{s}_{x}}\text{ and }a=\stackrel{¯}{y}-b\stackrel{¯}{x}$$

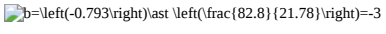
As before, the equation of the linear regression line is

$$\text{Predicted } y = a + b * x$$

Example: Highway Sign Visibility

We will now find the equation of the least-squares regression line using the output from a statistics package.

```
> summary(data)
   Age             Distance
   Min.   :18      Min.   :280
   1st Qu.:21.8    1st Qu.:82.8
   Median :54      Median :420
   Mean   :51      Mean   :423
   3rd Qu.:71.3    3rd Qu.:467.5
   Max    :82      Max    :590
> cor(data$Age,data$Distance)
   [1] -0.793
```

- The **slope** of the line is 
$$b=\left(-0.793\right)\ast \left(\frac{82.8}{21.78}\right)=-3$$
- The **intercept** of the line is a = 423 – (-3 * 51) = 576 and therefore the **least-squares regression line** for this example is Predicted distance = 576 + (-3 * Age), which can also be written as Predicted distance = 576 – 3 * Age

Try It

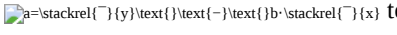
https://assessments.lumenlearning.co...sessments/3864

Try It

https://assessments.lumenlearning.co...sessments/3488

Now you know how to calculate the least-squares regression line from the correlation and the mean and standard deviation of x and y. But what do these formulas tell us about the least-squares line?

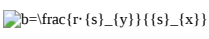
We know that the intercept a is the predicted value when x = 0.

The formula 
$$a=\stackrel{¯}{y}\text{}\text{}\text{−}\text{}b\cdot\stackrel{¯}{x}$$
tells us that the we can find the intercept using the point: (
$$\overline{x},\overline{y}$$
).

This is interesting because it says that every least-squares regression line contains this point. In other words, the least-squares regression line goes through the mean of x and the mean of y.

We also know that the slope of the least-squares regression line is the average change in the predicted response when the explanatory variable increases by 1 unit.

The slope formula



$$b=\frac{r\cdot{s}_{y}}{{s}_{x}}$$

tells us that the slope is related to the correlation in this way: when x increases an x standard deviation, the predicted y-value does not change by a y standard deviation. Instead, the predicted y-value changes by less than a y standard deviation. The change is a fraction of a y standard deviation, and that fraction is r. Another way to say this is that when x increases by a standard deviation in x, the average change in the predicted response is a fractional change of r standard deviations in y.

It is not surprising that slope and correlation are connected. We already know that when a linear relationship is positive, the correlation and the slope are positive. Similarly, when a linear relationship is negative, the correlation and slope are both negative.

But now we understand this connection more precisely.

Let's Summarize

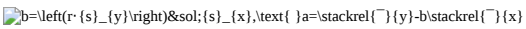
- The line that best summarizes a linear relationship is the least-squares regression line. The least-squares line is the best fit for the data because it gives the best predictions with the least amount of overall error. The most common measurement of overall error is the sum of the squares of the errors (SSE). The least-squares line is the line with the smallest SSE.
- We use the least-squares regression line to predict the value of the response variable from a value of the explanatory variable.
- Prediction for values of the explanatory variable that fall outside the range of the data is called extrapolation. These predictions are unreliable because we do not know if the pattern observed in the data continues outside the range of the data. Avoid making predictions outside the range of the data.
- The slope of the least-squares regression line is the average change in the predicted values of the response variable when the explanatory variable increases by 1 unit.
- We have two methods for finding the equation of the least-squares regression line:

$$\text{Predicted } y = a + b * x$$

Method 1: We use technology to find the equation of the least-squares regression line:

$$\text{Predicted } y = a + b * x$$

Method 2: We use summary statistics for x and y and the correlation. In this method we can calculate the slope b and the y-intercept a using the following:



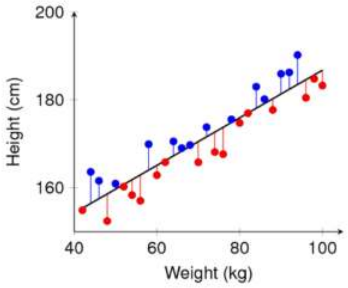
Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.3: Introduction to Assessing the Fit of a Line

What you'll learn to do: Use residuals, standard error, and r^2 to assess the fit of a linear model.



Graphing the regression line with the scatterplot gives a visual depiction of how well the regression line fits the data. To further hone in on assessing the fit of our regression line to the data, in this section we present:

- Residual plots.
- The correlation coefficient r gives us a numerical way to measure this fit.
- Interpreting the square of the correlation coefficient r^2 .
- Interpreting the standard error s_e.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.4: Assessing the Fit of a Line (1 of 4)

Learning Objectives

- Use residuals, standard error, and r^2 to assess the fit of a linear model.

Introduction

Let's take a moment to summarize what we have done up to this point in *Examining Relationships: Quantitative Data*. Our goal from the beginning was to *examine the relationship between two quantitative variables*. We started by looking at scatterplots to see if we could see any pattern between the explanatory and response variables. We focused early in the course on identifying those cases that were *linear* in form. At the same time, we assessed how strong the linear relationship was on the basis of visual inspection. As is our usual strategy, we turned from graphs to numeric measures, and in particular, we developed the correlation coefficient, *r*, as a measure of the strength of the linear relationship we observed in the graph.

Once we established that there was a linear relationship between explanatory and response variables, the next step was to find a line that fit the data: the *best-fit line*. Here we used the least-squares method to find the regression line. Finally, we used the equation of the regression line to predict the value of the response variable for a given value of the explanatory variable.

How Good Is the Best-Fit Line?

Now that we have a mathematical model (the least-squares regression line) that we can use to make predictions, we want to know: How good are these predictions, and how can we measure the error in a prediction?

Example

Highway Sign Visibility

Let's begin our investigation by predicting the maximum distance that an 18-year-old driver can read a highway sign and then determining the error in our prediction.

We use the regression line equation:

$$\text{Distance} = 576 + (-3 * \text{Age})$$

To predict the distance for an 18-year-old driver, we plug Age = 18 into the equation.

$$\text{Predicted distance} = 576 + (-3 * 18) = 522$$

Our prediction is that 522 feet is the maximum distance at which an 18-year-old driver can read a highway sign. Now let's compare our prediction to the actual data for the 18-year-old driver: (18, 510).

$$\text{The error in our prediction is } 510 - 522 = -12.$$

This tells us that the actual distance for the 18-year-old driver is 12 feet closer than the prediction. In other words, our prediction is too large. It overestimates the actual distance by 12 feet.

So in general, we have Observed data value – Predicted value = Error.

If we use (x, y) to represent a typical data point and $\hat{y}$ to represent the predicted value (obtained by using the regression equation), then we have

$$\begin{array}{l}\text{observed}-\text{predicted}=\text{error}\\ y-\hat{y}=\text{error}\end{array}$$

Try It

Using this table showing "observed" and "predicted" distances for some drivers, find the following:

	Age	Distance (observed)	Distance (predicted)	Error observed - predicted
Driver 1	18	510	576+(-3)(18) = 522	-12
Driver 2	32	410	576+(-3)(32) = 480	-70
Driver 3	55	420	576+(-3)(55) = 411	9
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Driver 30	82	360	.	.

https://assessments.lumenlearning.co...sessments/3497

https://assessments.lumenlearning.co...sessments/3498

https://assessments.lumenlearning.co...sessments/3499

Now let's look at the error from a different perspective. We can think of the error as a way to adjust the prediction to match the data value.

From this point of view, we rewrite $y-\hat{y}=\text{error}$ as $y=\hat{y}+\text{error}$.

This last equation says that the observed value is the predicted value plus the error. In other words, we can think of the error as the amount that we have to add to the prediction to get the observed value. From this point of view, the error can be thought of as a *correction term*. If the error is positive, it means the prediction is too small (the prediction underestimates the actual *y*-value). If the error is negative, it means the prediction is too large (the prediction overestimates the actual *y*-value).

The prediction error is also called a **residual**. So another way to express the previous equation is

$$y=\hat{y}+\text{residual}$$

In our next example, we look at prediction error from this point of view.

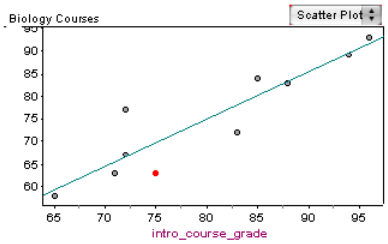
Example

Biology Courses

A biology department tracks the progress of students in its program. Grades in the introductory biology course have a strong linear relationship with grades in the upper-level biology courses ($r = 0.91$).

The least-squares regression equation is

$$\text{Upper course grade} = -8.9 + (1.05 * \text{Intro course grade})$$



	Intro grade	Upper grade
Student 1	65	58
Student 2	71	63
Student 3	72	67
Student 4	72	77
Student 5	75	63
Student 6	83	72
Student 7	85	84
Student 8	88	83
Student 9	94	89
Student 10	96	93

Let's look at the predicted upper course grade for a student who makes a 75% in the introductory biology course.

$$\text{Upper course grade} = -8.9 + (1.05 * 75) = 69.85 \approx 70$$

The regression line predicts that this student will make a 70% in the upper-level biology course.

The actual grade in the upper-level course for this student is 63%. The prediction is too high: it overestimates the data. To match the data value, we would need to subtract 7 from the prediction, so the error is −7.

In the scatterplot, notice that the regression line lies above the point (75, 63). Visually, we can see that the prediction is too high. This reinforces our previous observation that the prediction overestimates the data value. We would have to adjust the prediction downward to match the data value. Viewing the error as a correction term, we see the correction has to be negative.

Notice that when a point is close to the regression line, the prediction is close to the actual upper course grade, so the error is small. Another way to say this is that points close to the regression line have a small residual.

Try It

https://assessments.lumenlearning.co...sessments/3500

https://assessments.lumenlearning.co...sessments/3501

https://assessments.lumenlearning.co...sessments/3502

https://assessments.lumenlearning.co...sessments/3503

https://assessments.lumenlearning.co...sessments/3504

https://assessments.lumenlearning.co...sessments/3505

https://assessments.lumenlearning.co...sessments/3506

https://assessments.lumenlearning.co...sessments/3507

Contributors and Attributions

3.5: Assessing the Fit of a Line (2 of 4)

Learning Objectives

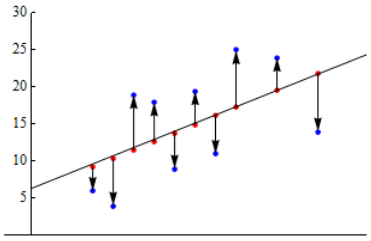
- Use residuals, standard error, and r^2 to assess the fit of a linear model.

Introduction

Now we move from calculating the residual for an individual data point to creating a graph of the residuals for all the data points. We use residual plots to determine if the linear model fits the data well.

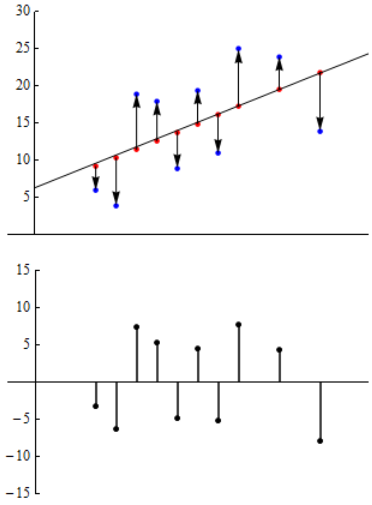
Residual Plots

The graph below shows a scatterplot and the regression line for a set of 10 points. The blue points represent our original data set, that is, our observed values. The red points, lying directly on the regression line, are the predicted values.



The vertical arrows from the predicted to observed values represent the residuals. The up arrows correspond to positive residuals, and the down arrows correspond to negative residuals.

Now consider the following pair of graphs. The top graph is a copy of the graph we looked at above. In the graph below, we plotted the values of the residuals on their own. (The explanatory variable is still plotted on the horizontal axis, though it is not indicated this here.) This is called a **residual plot**.



In the residual plot, each point with a value greater than zero corresponds to a data point in the original data set where the observed value is greater than the predicted value. Similarly, negative values correspond to data points where the observed value is less than the predicted value.

What are we looking for in a residual plot?

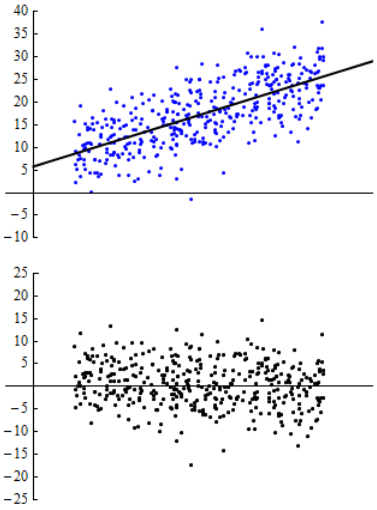
We use residual plots to determine if a linear model is appropriate. In particular, we look for any *unexpected patterns* in the residuals that may suggest that the data is not linear in form.

To help us identify an unexpected pattern, we start by looking at what we *expect* to see in a residual plot *when the form is linear.*

Example

No Pattern in Residual Plot

Consider the pair of graphs below. Here we have a scatterplot for a data set consisting of 400 observations. The regression line is shown in the scatterplot. The residual plot is below the scatterplot.



In this example, the line in the scatterplot is a good summary of the positive linear pattern in the data. Notice that the points in the residual plot seem to be randomly scattered. As we examine the residuals from left to right, they don't appear to follow a particular path, nor does the cloud of points widen or narrow in any systematic way. We see no particular pattern. Thus, in the ideal case, when a linear model is really a good fit, we expect to see *no pattern* in the residual plot.

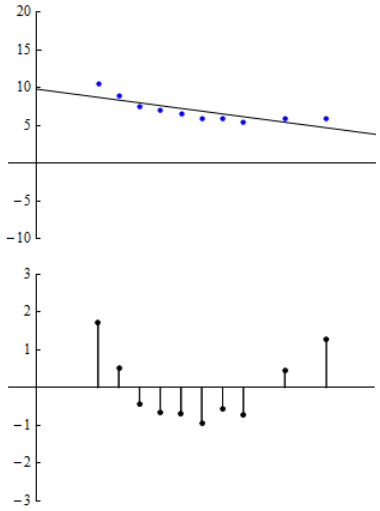
Our general principle when looking at residual plots, then, is that a residual plot with *no* pattern is good because it suggests that our use of a linear model is appropriate.

However, we must be flexible in applying this principle because what we see usually lies somewhere between the extremes of no pattern and a clear pattern. Let's look at some specific examples.

Example

Patterns in Residual Plots

At first glance, the scatterplot appears to show a strong linear relationship. The correlation is $r = 0.84$. However, when we examine the residual plot, we see a clear U-shaped pattern. Looking back at the scatterplot, this movement of the data points above, below and then above the regression line is noticeable. The residual plot, particularly when graphed at a finer scale, helps us to focus on this deviation from linearity.

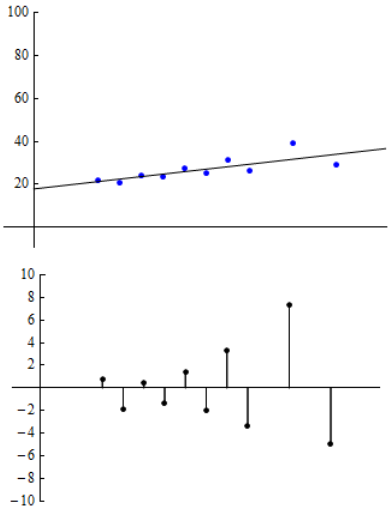


The pattern in the residual plot suggests that our linear model may not be appropriate because the model predictions will be too high for values in the middle of the range of the explatory variable and too low for values at the two ends of that range. A model with a curvilinear form may be more appropriate.

Example

Patterns in Residual Plots 2

This scatterplot is based on datapoints that have a correlation of $r = 0.75$. In the residual plot, we see that residuals grow steadily larger in absolute value as we move from left to right. In other words, as we move from left to right, the observed values deviate more and more from the predicted values. Again, we have chosen a smaller vertical scale for the residual plot to help amplify the pattern to make it easier to see.

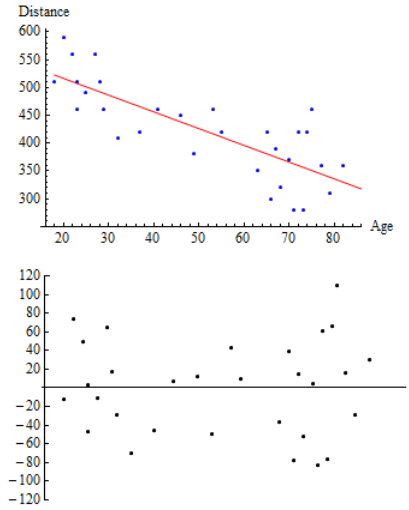


The pattern in the residual plot suggests that predictions based on the linear regression line will result in greater error as we move from left to right through the range of the explatory variable.

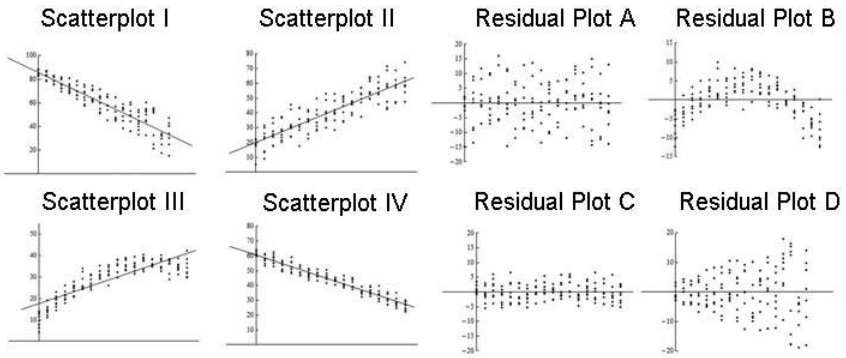
Example

Highway Sign Visibility

Let’s return now to our original example and take a look at what the residual plot tell us about the appropriateness of applying a linear model to this data.



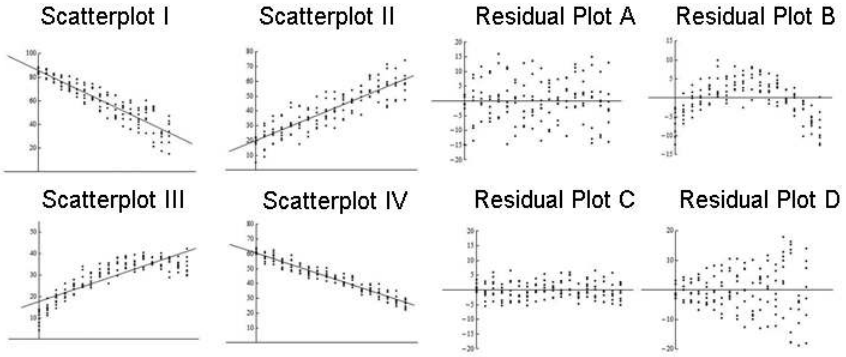
Note that the residuals are fairly randomly dispersed. However, they seem to be a bit more spread out on the left and right than they are in the middle. As we look at higher ages, there seems to be greater variation in the residuals, which suggests that we may want to be more cautious if we are trying to predict distances for older drivers. And the risks associated with extrapolation beyond the range of the data seem to be even greater here. In this case, we may still use this linear model but condition the use of it on our analysis of the residual plot.



Try It

https://assessments.lumenlearning.co...sessments/3508

Here again are four scatterplots with regression lines shown and four corresponding residual plots.



Try It

https://assessments.lumenlearning.co...sessments/3509

https://assessments.lumenlearning.co...sessments/3510

https://assessments.lumenlearning.co...sessments/3511

https://assessments.lumenlearning.co...sessments/3512

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.6: Assessing the Fit of a Line (3 of 4)

Learning Objectives

- Use residuals, standard error, and r^2 to assess the fit of a linear model.

Introduction

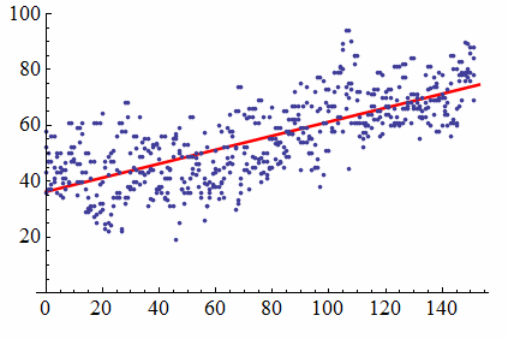
Here we continue our discussion of the question, *How good is the best-fit line?*

Let's summarize what we have done so far to address this question. We began by looking at how the predictions from the least-squares regression line compare to observed data. We defined a residual to be the amount of error in a prediction. Next, we created residual plots. A residual plot with no pattern reassures us that our linear model is a good summary of the data.

But how do we know if the explanatory variable we chose is really the best predictor of the response variable?

The regression line does not take into account other variables that might also be good predictors. So let's investigate the question, *What proportion of the variation in the response variable does our regression line explain?*

We begin our investigation with a scatterplot of the daily high temperature (°F) in New York City from January 1 to June 1. We have 4 years of data (2002, 2003, 2005, and 2006). The least-squares regression line has the equation $y = 36.29 + 0.25x$, where x is the number of days after January 1. Therefore, January 1 corresponds to $x = 0$, and June 1 corresponds to $x = 151$.



Two things stand out as we look at this picture. First, we see a clear, positive linear relationship tracked by the regression line. As the days progress, there is an associated increase in temperature. Second, we see a substantial scattering of points around the regression line. We are looking at 4 years of data, and we see a lot of variation in temperature, so the day of the year only partially explains the increase in temperature. Other variables also influence the temperature, but the line accounts only for the relationship between the day of the year and temperature.

Now we ask the question, *Given the natural variation in temperature, what proportion of that variation does our linear model explain?*

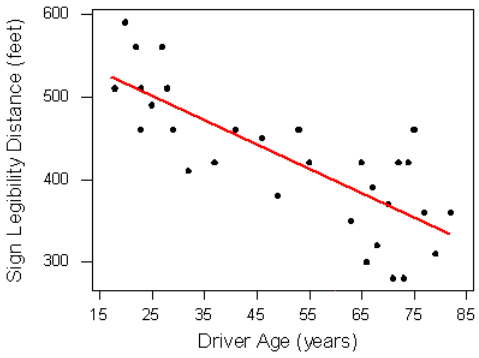
The answer, which is surprisingly easy to calculate, is just the square of the correlation coefficient.

The value of r^2 is the proportion of the variation in the response variable that is explained by the least-squares regression line.

In the present case, we have $r = 0.73$; therefore, $\frac{\text{explained variation}}{\text{total variation}}={0.73}^{2}=0.53$. And so we say that our linear regression model explains 53% of the total variation in the response variable. Consequently, 47% of the total variation remains unexplained.

Example

Highway Sign Visibility



Recall that the least-squares regression line is Distance = 576 − 3 * Age. The correlation coefficient for the highway sign data set is −0.793, so $r^2 = (-0.793)^2 = 0.63$.

Our linear model uses age to predict maximum distance at which a driver can read a highway sign. Other variables may also influence reading distance. We can say the linear relationship between age and maximum reading distance accounts for 63% of the variation in maximum reading distance.

Try It

https://assessments.lumenlearning.co...sessments/3513

https://assessments.lumenlearning.co...sessments/3868

Try It

https://assessments.lumenlearning.co...sessments/3514

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.7: Assessing the Fit of a Line (4 of 4)

Learning Objectives

- Use residuals, standard error, and r^2 to assess the fit of a linear model.

Introduction

Our final investigation into assessing the fit of the regression line focuses on typical error in the predictions.

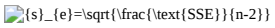
Previously, we calculated the error in a single prediction by calculating

$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$

But we use the regression line to make predictions even when we do not have an observed value, so we need a method for using all of the residuals to compute a typical amount of error.

We ask the question, *How do we measure the typical amount of error for predictions from the regression line?*

The most common measure of the size of the typical error is the **standard error of the regression**, which is represented by s_e. It is calculated using the following formula:

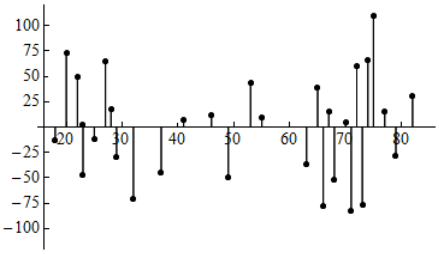


where *SSE* stands for the sum of the squared errors.

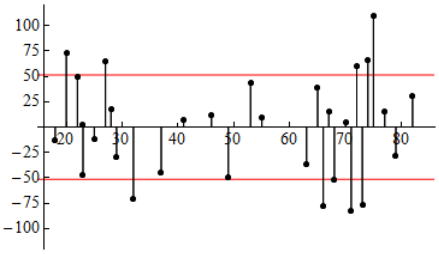
Finding the standard error of the regression is similar to finding the standard deviation of a distribution of data points from a single quantitative variable. In *Summarizing Data Graphically and Numerically*, we learned that the *standard deviation is roughly a measure of average distance about the mean*. Here the *standard error is roughly a measure of the average distance of the points about the regression line*.

Let's return to our example where age is used to predict the maximum distance for reading highway signs.

The residual plot for the highway sign data set is shown below. We can visualize the SSE in the formula as simply the sum of the squares of all of the vertical (residual) line segments. After dividing by $n - 2$, we have the average *squared* residual. Taking the square root then gives us a measure of the average size of the residuals.



In the case of the highway sign data, the value of s_e is 51.35. In the figure below, we added horizontal lines at $y = 51.35$ and $y = -51.35$, so the red line represents the typical size of the error.



Comment: When we mark the s_e on this residual plot, errors that fall outside of this range are larger than average. We see again that most of the errors that exceed ±51.35 are on the right. This illustrates that predictions of maximum reading distance for older drivers have larger error.

Note: Most statistics software computes r and r^2 and s_e. Therefore, our focus is not on calculating but on understanding and interpreting.

Now let's apply the standard error of the regression as a measurement of typical error.

Example

Highway Sign Visibility

Let's take another look at the prediction we made earlier using the regression line equation:

$$\text{Distance} = 576 + (-3 * \text{Age})$$

In a previous example, we predicted the maximum distance that a 60-year-old driver can read a highway sign. We plugged Age = 60 into the equation and found that

$$\text{Predicted distance} = 576 + (-3 * 60) = 396$$

The question we now ask is, How good is this prediction?

Unfortunately, there is no 60-year-old driver in the original data set of 30 drivers, so we cannot calculate the residual. Instead, we use the s_e as a measurement of typical error.

Technology gives $s_e = 51.35$.

So how good is the prediction for the 60-year-old driver? Based on the s_e for this data, we estimate that our prediction of 396 feet is off by ±51 feet.

	Intro grade(%)	Upper grade(%)	Predictions	Error (Residual)	Error Squared
Student 1	65	58	59.1	−1.1	1.21
Student 2	71	63	65.4	−2.4	5.76
Student 3	72	67	66.4	0.6	0.36
Student 4	72	77	66.4	10.6	112.36
Student 5	75	63	69.6	−6.6	43.56
Student 6	83	72	77.9	−5.9	34.81
Student 7	85	84	80	4	16
Student 8	88	83	83.2	−0.2	0.04
Student 9	94	89	89.5	−0.5	0.25
Student 10	96	93	91.5	1.5	2.25

Try It

https://assessments.lumenlearning.co...sessments/3515

https://assessments.lumenlearning.co...sessments/3516

https://assessments.lumenlearning.co...sessments/3517

Try It

https://assessments.lumenlearning.co...sessments/3869

Let's Summarize

- When we use a regression line to make predictions, there is error in the prediction. We calculate this error as **Observed data value − Predicted value**. A residual is another name for the prediction error.
- We use residual plots to determine whether a linear model is a good summary of the relationship between the explanatory and response variables. In particular, we look for any *unexpected patterns* in the residuals that may suggest the data is not linear in

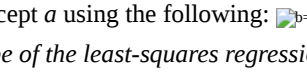
form.

- We have two numeric measures to help us judge how well the regression line models the data.
  - The square of the correlation coefficient, r^2, is the proportion of the variation in the response variable that is explained by the least-squares regression line.
  - The standard error of the regression, s_e, gives a typical prediction error based on all of the data. It roughly measures the average distance of the data from the regression line. In this way, it is similar to the standard deviation, which roughly measures average distance from the mean.

Contributors and Attributions

3.8: Putting It Together- Examining Relationships- Quantitative Data

Let's Summarize

- We use a *scatterplot* to graph the relationship between two quantitative variables. In a scatterplot, each dot represents an individual. We always plot the explanatory variable on the horizontal x-axis.
- When we explore a relationship between two quantitative variables using a scatterplot, we describe the overall pattern (*direction*, *form*, and *strength*) and deviations from the pattern (*outliers*).
- When the *form of a relationship is linear*, we use the correlation coefficient, *r*, to measure the strength and direction of the linear relationship. The correlation ranges between −1 an 1. If the pattern is linear, an *r*-value near −1 indicates a strong negative linear relationship and an *r*-value near +1 indicates a strong positive linear relationship. Following are some cautions about interpreting correlation:
  - **Always make a scatterplot before interpreting *r*.** Correlation is affected by outliers and should be used only when the pattern in the data is linear.
  - **Association does not imply causation.** Do not interpret a high correlation between explanatory and response variables as a cause-and-effect relationship.
  - **Beware of lurking variables** that may be explaining the relationship seen in the data.
- The line that best summarizes a linear relationship is the *least-squares regression line*. The least-squares line is the best fit for the data because it gives the best predictions with the least amount of error. The most common measurement of overall error is the sum of the squares of the errors, SSE. The least-squares line is the line with the smallest SSE.
- We use the least-squares regression line to predict the value of the response variable from a value of the explanatory variable. **Avoid making predictions outside the range of the data.** (This is called *extrapolation*.)
- We have two methods for *finding the equation of the least-squares regression line:* Predicted $y = a + b * x$
  - We use technology to find the equation of the least-squares regression line: Predicted $y = a + b * x$
  - We use summary statistics for x and y and the correlation. Using this method, we can calculate the slope b and the y-intercept a using the following: 
- The *slope of the least-squares regression* line is the average change in the predicted values when the explanatory variable increases by 1 unit.
- When we use a regression line to make predictions, there is error in the prediction. We calculate this error as Observed value − Predicted value. This prediction error is also called a *residual*.
- We use *residual plots* to determine whether a linear model is a good summary of the relationship between the explanatory and response variables. In particular, we look for any unexpected patterns in the residuals that may suggest that the data is not linear in form.
- We have two numeric measures to help us judge how well the regression line models the data:
  - The square of the correlation, r^2 , is the proportion of the variation in the response variable that is explained by the least-squares regression line.
  - The standard error of the regression, s_e , gives a typical prediction error based on all of the data. It roughly measures the average distance of the data from the regression line. In this way, it is similar to the standard deviation, which roughly measures average distance from the mean.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.9: StatTutor- Academic Performance

You are now ready to practice what you learned in this module by doing a StatTutor exercise. StatTutor exercises are designed to help you apply what you have learned to a real life data analysis question.

Instructions: One of the first few screens in StatTutor has a link to download the dataset for this StatTutor exercise. When you click that link, a pop-up window will appear asking if you want to open or save the file. Make sure you click “Save,” which will allow you to save the file to your hard drive. Then find the downloaded file and double-click it to open it if you’re using R, Minitab, Excel, or StatCrunch, or transfer it to your calculator if you’re using the TI Calculator.

A link to an interactive elements can be found at the bottom of this page.

Contributors and Attributions

3.10: Assignment- Scatterplot

In this exercise we will:

- Learn how to create a scatterplot.
- Use the scatterplot to examine the relationship between two quantitative variables.
- Learn how to create a labeled scatterplot.
- Use the labeled scatterplot to better understand the form of a relationship.

In this activity we explore the relationship between weight and height for 81 adults. We will use height as the explanatory variable. Weight is the response variable.

We will then label the men and women by adding the categorical variable gender to the scatterplot. We will see if separating the groups contributes to our understanding of the form of the relationship between height and weight.

Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel | TI Calculator

Question 1:

Describe the relationship between the height and weight of the subjects. To describe the relationship write about the pattern (direction, form, and strength) and any deviations from the pattern (outliers).

So far we have studied the relationship between height and weight for all of the males and females together. It may be interesting to examine whether the relationship between height and weight is different for males and females. To visualize the effect of the third variable, gender, we will indicate in the scatterplot which observations are males and which are females.

Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel | TI Calculator

Question 2:

Compare and contrast the relationship between height and weight for males and females. To compare and contrast the relationships by gender write about the pattern (direction, form, and strength) and any deviations from the pattern (outliers) for each group.

Discuss how the patterns for the two groups are similar and how they are different.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.11: Assignment- Linear Relationships

In this activity we will:

- Learn how to compute the correlation.
- Practice interpreting the value of the correlation.
- See an example of how including an outlier can *increase* the correlation.

Recall the following example: The average gestation period, or time of pregnancy, of an animal is closely related to its longevity—the length of its lifespan. Data on the average gestation period and longevity (in captivity) of 40 different species of animals have been recorded.

Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel | TI Calculator

Remember that the correlation is only an appropriate measure of the **linear** relationship between two quantitative variables. First produce a scatterplot to verify that gestation and longevity are nearly linear in their relationship.

Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel | TI Calculator

Observe that the relationship between gestation period and longevity is linear and positive. Now we will compute the correlation between gestation period and longevity.

Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel | TI Calculator

Question 1:

Report the correlation between gestation and longevity and comment on the strength and direction of the relationship. Interpret your findings in context.

Now return to the scatterplot that you created earlier. Notice that there is an outlier in both longevity (40 years) and gestation (645 days). Note: This outlier corresponds to the longevity and gestation period of the elephant.

What do you think will happen to the correlation if we remove this outlier?

Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

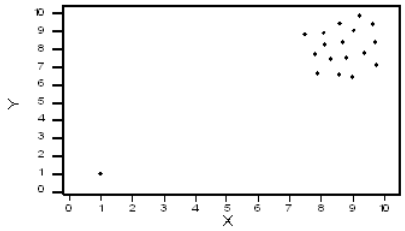
R | StatCrunch | Minitab | Excel | TI Calculator

Question 2:

Report the new value for the correlation between gestation and longevity and compare it to the value you found earlier when the outlier was included. What is it about this outlier that results in the fact that its inclusion in the data causes the correlation to increase? (Hint: look at the scatterplot.)

Comment

In the last activity, we saw an example where there was a positive linear relationship between the two variables, and including the outlier just “strengthened” it. Consider the hypothetical data displayed by the following scatterplot:



In this case, the low outlier gives an “illusion” of a positive linear relationship, whereas in reality, there is no linear relationship between X and Y.

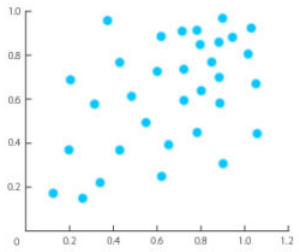
Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.12: Introduction to Scatterplots

What you'll learn to do: Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.



When investigating relationships between two quantitative variables, scatterplots are a simple way to visually represent the spread, direction, strength of relationship, and potential outliers of the data. With larger datasets, a scatterplot can more succinctly display the overall pattern than when the data are presented as a table. This visualization can also hint at the general shape of the relationship (for example, increasing linear, decreasing linear, or non-linear curves) while also helping us identify any deviations from that pattern.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.13: Assignment- Linear Regression

In this activity we will:

- Find a regression line and plot it on the scatterplot.
- Examine the effect of outliers on the regression line.
- Use the regression line to make predictions and evaluate how reliable these predictions are.

Background

The modern Olympic Games have changed dramatically since their inception in 1896. For example, many commentators have remarked on the change in the quality of athletic performances from year to year. Regression will allow us to investigate the change in winning times for one event—the 1,500 meter race.

Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel | TI Calculator

Observe that the form of the relationship between the 1,500 meter race's winning time and the year is linear. The least squares regression line is therefore an appropriate way to summarize the relationship and examine the change in winning times over the course of the last century. We will now find the least squares regression line and plot it on a scatterplot.

Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel | TI Calculator

Question 1:

Give the equation for the least squares regression line, and interpret it in context.

Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel | TI Calculator

Question 2:

Give the equation for this new line and compare it with the line you found for the whole dataset, commenting on the effect of the outlier.

Question 3:

Our least squares regression line associates years as an explanatory variable, with times in the 1,500 meter race as the response variable. Use the least squares regression line you found in question 2 to predict the 1,500 meter time in the 2008 Olympic Games in Beijing. Comment on your prediction.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.14: Scatterplots (1 of 5)

Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

Example

Highway Signs

A research firm conducts a study to explore the relationship between a driver's age and the driver's ability to read highway signs. The subjects are a random sample of 30 drivers between the ages of 18 and 82. *(Source: Jessica M. Utts and Robert F. Heckard, Mind on Statistics [Brooks/Cole, 2002]. Original source: Data collected by The Last Resource, Inc., Bellfonte, PA.)*

Because the purpose of this study is to explore the effect of age on the driver's ability to read highway signs,

- the *explanatory* variable is *age*, and
- the *response* variable is the maximum distance at which the driver can read a highway sign, or *maximum reading distance*.

Both variables are quantitative.

Here is what the raw data look like:

Explanatory **Response**

	Age	Distance
Driver 1	**18**	**510**
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360

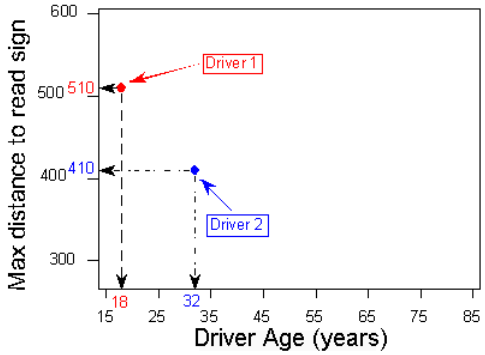
In this data set, the individuals are the 30 drivers. For each driver, we have two values: age and maximum reading distance.

To explore the relationship between age and distance, we create a graph called a **scatterplot**. To create a scatterplot, we use an ordered pair (x, y) to represent the data for each driver. The ***x*-coordinate** is the explanatory variable: driver's age. The ***y*-coordinate** is the response variable: maximum reading distance.

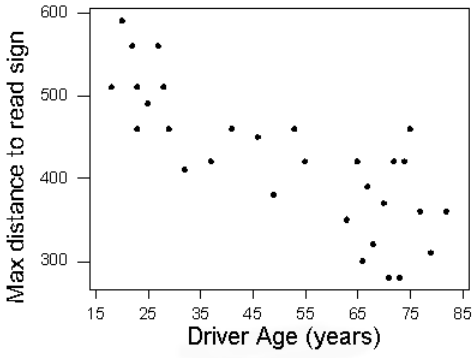
For this example, the ordered pair (18, 510) represents an 18-year-old driver who can read a highway sign at a maximum distance of 510 feet. We plot a point for each ordered pair. In the scatterplot, each driver appears as a single point.

Generally, each point in a scatterplot represents *one individual*. The x-coordinate is the value of the explanatory variable for that individual. The y-coordinate is the value of the response variable for that individual.

	Age (X)	Distance (Y)
Driver 1	**18**	**510**
Driver 2	**32**	**410**
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360

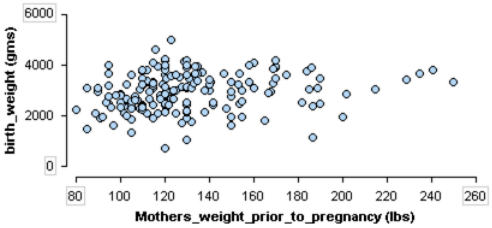


Here is the completed scatterplot:



Try It

Recall this dataset from a medical study. In this study researchers collected data on new mothers to identify variables connected to low birth weights. This scatterplot investigates the relationship between two quantitative variables in the study: mother's weight prior to pregnancy and baby's birth weight.



https://assessments.lumenlearning.co...sessments/3856

https://assessments.lumenlearning.co...sessments/3466

https://assessments.lumenlearning.co...sessments/3467

https://assessments.lumenlearning.co...sessments/3468

https://assessments.lumenlearning.co...sessments/3469

Comment

Remember: The explanatory variable is on the horizontal x-axis. The response variable is on the vertical y-axis. Sometimes the variables do not have a clear explanatory–response relationship. In this case, there is no rule to follow. Plot the variables on either axis.

Contributors and Attributions

3.15: Scatterplots (2 of 5)

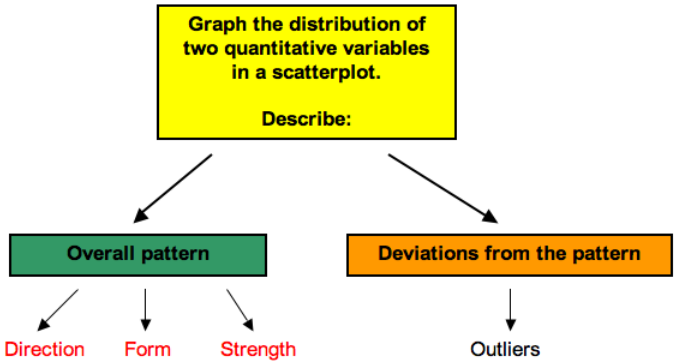
Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

Interpreting the Scatterplot

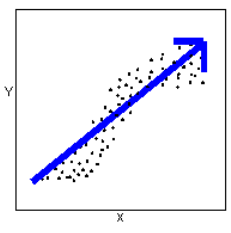
How do we describe the relationship between two quantitative variables using a scatterplot? We describe the overall pattern and deviations from that pattern.

This is the same way we described the distribution of one quantitative variable using a dotplot or a histogram in *Summarizing Data Graphically and Numerically*. To describe the overall pattern of the distribution of one quantitative variable, we describe the shape, center, and spread. We also describe deviations from the pattern (outliers).

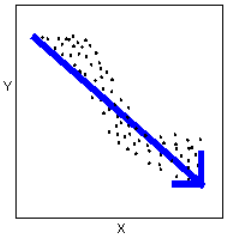


Similarly, in a scatterplot, we describe the overall pattern with descriptions of **direction**, **form**, and **strength**. Deviations from the pattern are still called outliers.

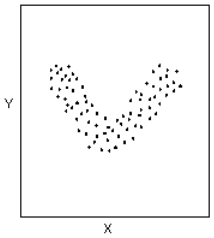
- The direction of the relationship can be positive, negative, or neither:



Positive relationship



Negative relationship



Neither positive nor negative

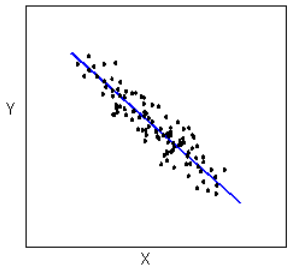
A *positive (or increasing) relationship* means that an increase in one of the variables is associated with an increase in the other.

A *negative (or decreasing) relationship* means that an increase in one of the variables is associated with a decrease in the other.

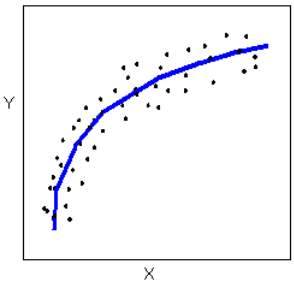
Not all relationships can be classified as either positive or negative.

- The form of the relationship is its general shape. To identify the form, describe the shape of the data in the scatterplot. In practice, forms that we commonly use have mathematical equations. We look at a few of these equations in this course. For now, we simply describe the shape of the pattern in the scatterplot. Here are a couple of forms that are quite common:**Linear**

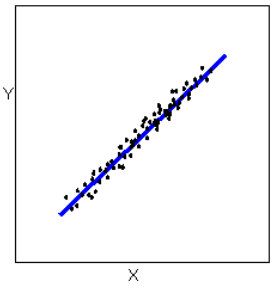
form: The data points appear scattered about a line. We use a line to summarize the pattern in the data. We study the equation for a line in this module.



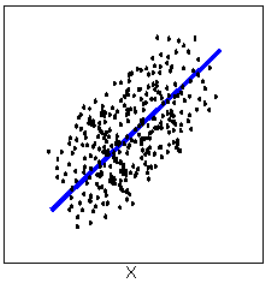
Curvilinear form: The data points appear scattered about a smooth curve. We use a curve to summarize the pattern in the data. We study some specific types of curvilinear forms with their equations in Modules 4 and 12.



- The strength of the relationship is a description of how closely the data follow the form of the relationship. Let's look, for example, at the following two scatterplots displaying positive, linear relationships:



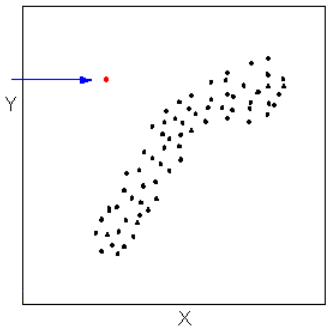
strong relationship



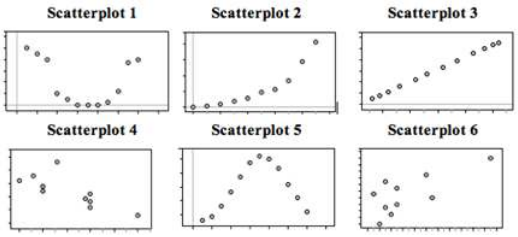
weaker relationship

In the top scatterplot, the data points closely follow the linear pattern. This is an example of a *strong linear* relationship. In the bottom scatterplot, the data points also follow a linear pattern, but the points are not as close to the line. The data is more scattered about the line. This is an example of a *weaker linear* relationship.Labeling a relationship as strong or weak is not very precise. We develop a more precise way to measure the strength of a relationship shortly.

Outliers are points that deviate from the pattern of the relationship. In the scatterplot below, there is one outlier.



Try It



Fill in the letter of the description that matches each scatterplot.

Descriptions:

A: X = month (January = 1), Y = rainfall (inches) in Napa, CA in 2010 (Note: Napa has rain in the winter months and months with little to no rainfall in summer.)

B: X = month (January = 1), Y = average temperature in Boston MA in 2010 (Note: Boston has cold winters and hot summers.)

C: X = year (in five-year increments from 1970), Y = Medicare costs (in $) (Note: the yearly increase in Medicare costs has gotten bigger and bigger over time.)

D: X = average temperature in Boston MA (°F), Y = average temperature in Boston MA (°C) each month in 2010

E: X = chest girth (cm), Y = shoulder girth (cm) for a sample of men

F: X = engine displacement (liters), Y = city miles per gallon for a sample of cars (Note: engine displacement is roughly a measure of engine size. Large engines use more gas.)

https://assessments.lumenlearning.co...sessments/3470

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

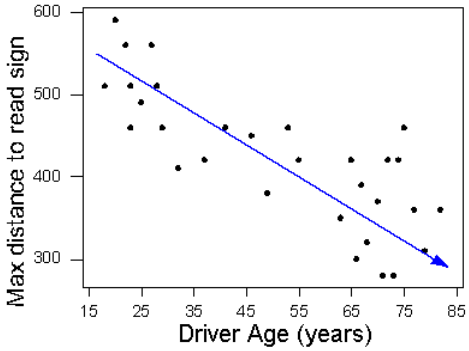
3.16: Scatterplots (3 of 5)

Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

Now we return to our previous example. We apply the ideas of direction, form, and strength to describe the relationship between the age of the driver and the maximum distance to read a highway sign. Here is the scatterplot:

Direction: The direction of the relationship is negative. An increase in age is associated with a decrease in reading distance, which makes sense because older drivers tend to have diminished eyesight. So most older drivers can read the sign only when they are close to it. In other words, they have a shorter maximum reading distance.



Form: The form of the relationship is linear.

Strength: The data points are fairly close to the line, so the relationship is moderately strong. Do not worry if you feel uncertain about describing the strength of a relationship. We mentioned earlier that descriptions of strength are not very precise. We develop a more precise measure of the strength shortly.

Outliers: There are no outliers. All the data points tend to follow the linear pattern.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.17: Scatterplots (4 of 5)

Learning Objectives

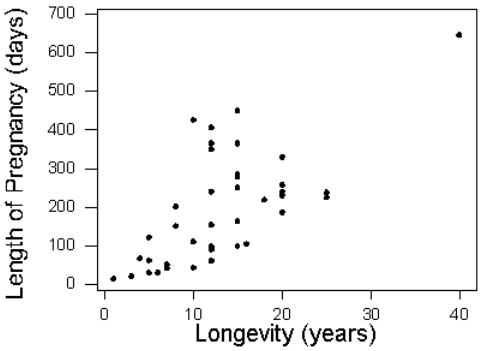
- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

We now look at two more examples:

Example

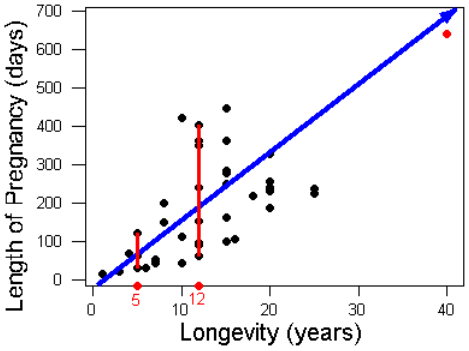
Average Length of Pregnancy

What is the relationship between an animal's lifespan and the length of its pregnancy? To investigate this question, we have data from 40 different species of animals living in captivity. We use average lifespan as the explanatory variable, *x*. The average length of pregnancy is the response variable, *y*. (Source: Allen J. Rossman and Beth L. Chance, *Workshop Statistics: Discovery with Data and Minitab* [Key College Publishing, 2001]. Original source: *World Almanac and Book of Facts, 1993* [World Almanac, 1993].)



What can we learn about the relationship from the scatterplot?

The *direction* of the relationship is positive. An increase in lifespan is associated with an increase in pregnancy length. In other words, animals that live longer tend to have longer pregnancies. The *form* of the relationship is linear. The relationship is moderately *strong*.



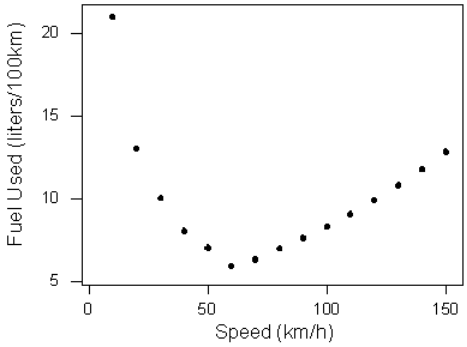
Is there an outlier? There is a data point that deviates from the rest of the data because it has large *x*– and *y*-values. This is the elephant. Elephants live a long time (large *x*-value) and have a long pregnancy (large *y*-value). So the elephant is an outlier in the distribution of both the lifespan and the pregnancy variables. But this data point follows the positive direction of the data and fits the linear pattern. With respect to the form and direction of the relationship between the variables, this point is not an outlier.

Notice that the variation in pregnancy length is larger for animals that live longer. For example, animals that live 5 years have pregnancies that range from about 30 days to 120 days. The short, red vertical line on the left illustrates this range. Animals that live 12 years have pregnancies that vary more, ranging from about 60 days to over 400 days. The longer red vertical line on the right illustrates this range. So the relationship is stronger for animals with shorter lifespans.

Example

Fuel Usage

When you drive a car, what is the relationship between the speed you drive and the amount of gas the car uses? In this study, engineers measured the amount of fuel (in liters) used to drive 100 kilometers. They made these fuel measurements for a car driving at a fixed speed (in kilometers per hour). They then made fuel measurements for different fixed speeds.



What can we learn about the relationship from the scatterplot?

The data describe a relationship that decreases, then increases, so the direction of the relationship is negative and then becomes positive. In other words, at slow speeds, the car uses a lot of fuel. The amount of fuel decreases rapidly to a low point when the speed is 60 kilometers per hour, so the car uses the least amount of fuel at a speed of 60 km/h. The amount of fuel increases gradually for speeds above 60 km/h. This forms a *curvilinear* relationship that is very *strong*. All of the data fit a smooth curve.

Is there an outlier? The point (10, 21) lies above the rest of the data. With respect to speed (x), this point is not an outlier. The x-value does not deviate from the pattern for the other x-values in the data. In this study, it appears that the engineers varied the speeds by increments of 10 km/h. However, the y-value is much higher than the other y-values. With respect to fuel usage, this point is an outlier. But the point fits the overall curvilinear pattern in the data, so with respect to direction and form, this point is not an outlier.

Try It

https://assessments.lumenlearning.co...sessments/3476

Comment

In *Summarizing Data Graphically and Numerically*, we developed a method for identifying outliers in a distribution of one quantitative variable. The method was the 1.5 * IQR rule. In a scatterplot, you can use this rule to determine if the x-value of a point is an outlier with respect to the x-values in data. Similarly, you can use this rule to determine if a y-value of a point is an outlier with respect to the y-values in the data. However, this rule does not help us identify a point that deviates from the overall pattern in the data.

Is there a method to identify outliers that deviate from the overall pattern in a scatterplot? The answer is yes, but we do not discuss these techniques in this course. For now, just look at the scatterplot and see if a point deviates from the overall pattern. In other words, see if the point deviates from the direction and form of the data. We will see later that this type of outlier can influence measures of center and spread for two quantitative variables.

Contributors and Attributions

3.18: Scatterplots (5 of 5)

Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

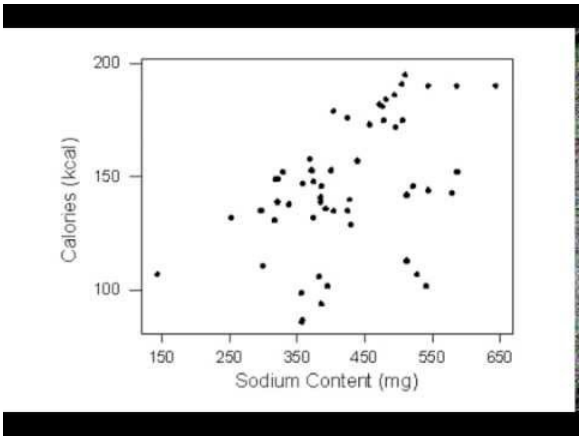
Labeling Groups in a Scatterplot

If we graph data from two or more groups in a scatterplot, the relationship between the two quantitative variables can be hidden or unclear. We can use a categorical variable to label groups within the scatterplot, then look for patterns within each group. The relationship may be clearer within each group.

Example

Hot Dogs

A study was conducted by a concerned health group in which 54 major hot dog brands were examined. Using this data, we explore the relationship between sodium content and calories. We begin by making a scatterplot with data from the three types of hot dogs: beef, poultry, and meat (meat is a combination of pork, beef, and poultry).



A YouTube element has been excluded from this version of the text. You can view it online here: pb.libretexts.org/cis/?p=148

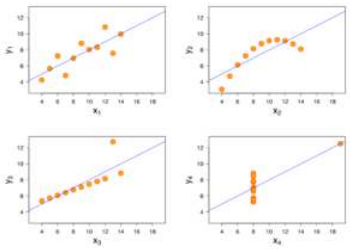
Let's Summarize

- The relationship between two quantitative variables is visually displayed using the scatterplot, where each point represents an individual. We always plot the explanatory variable on the horizontal x-axis and the response variable on the vertical y-axis.
- When we explore a relationship using the scatterplot, we should describe the *overall pattern* of the relationship and any *deviations* from that pattern. To describe the overall pattern, consider the *direction*, *form*, and *strength* of the relationship. Assessing the strength just by looking at the scatterplot can be problematic; using a numerical measure to determine strength is discussed later in this course.
- Adding labels to the scatterplot that indicate different groups or categories within the data might help us gain more insight about the relationship we are exploring.

Contributors and Attributions

3.19: Introduction to Linear Relationships

What you'll learn to do: Use a correlation coefficient to describe the direction and strength of a linear relationship. Recognize its limitations as a measure of the relationship between two quantitative variables.



Scatterplots are an excellent way to visually inspect the data, but to further investigate the relationship, it would help to quantify some metrics about the relationship. In particular, we are interested in:

- Direction: Does the response variable increase with the dependent variable? Or does the response variable decrease with the dependent variable?
- Strength: Does the scatterplot cluster tightly around a line?
- Form: Is the scatterplot evenly clustered around the line or are there regions where the scatter is more spread out? Does the shape of the scatterplot seem linear or curvilinear?

Contributors and Attributions

3.20: Linear Relationships (1 of 4)

Learning Objectives

- Use a correlation coefficient to describe the direction and strength of a linear relationship. Recognize its limitations as a measure of the relationship between two quantitative variables.

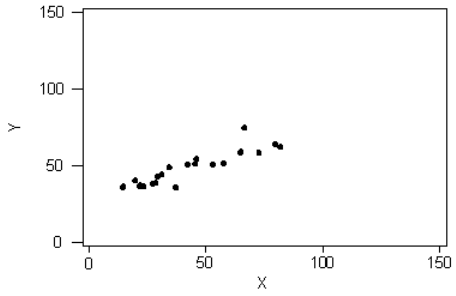
Introduction

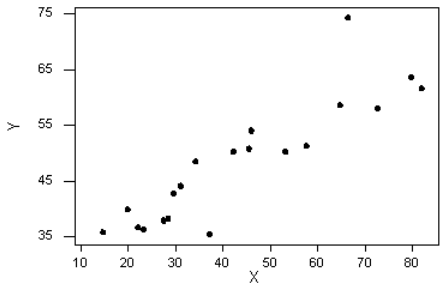
So far, we have visualized relationships between two quantitative variables using scatterplots. We have also described the overall pattern of a relationship by considering its direction, form, and strength. We noted that it is difficult to assess the strength of a relationship just by looking at the scatterplot. In this section, we develop a numerical measure to assess the strength.

We focus only on relationships that have a linear form. Linear forms are quite common and relatively simple to detect. More important, we have a numerical measure that can assess the strength of the linear relationship. We use this measure along with the scatterplot to describe the linear relationship between two quantitative variables.

Even though we now focus only on linear relationships, remember that not every relationship between two quantitative variables has a linear form. We have already seen several examples of relationships that are not linear. However, the measure of strength that we are about to study can be used only with linear relationships. If we use this measure with nonlinear relationships, we will draw incorrect conclusions about the relationship between the variables.

Let's start with an example. Consider the following two scatterplots.





We can see that in both cases, the direction of the relationship is *positive* and the form of the relationship is *linear*. What about the strength? Recall that the strength of a relationship is a description of how closely the data follow its form.

Try It

https://assessments.lumenlearning.co...sessments/3465

The scale used in a scatterplot can sometimes affect our assessment of strength, so we need to develop a measure for the strength of a linear relationship between two quantitative variables.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.21: Linear Relationships (2 of 4)

Learning Objectives

- Use a correlation coefficient to describe the direction and strength of a linear relationship. Recognize its limitations as a measure of the relationship between two quantitative variables.

The Correlation Coefficient (*r*)

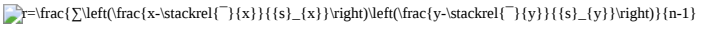
The numerical measure that assesses the strength of a linear relationship is called the **correlation coefficient** and is denoted by ***r***. In this section, we

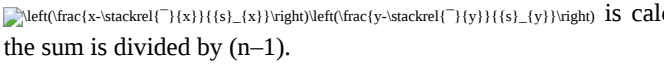
- define *r*.
- discuss the calculation of *r*.
- explain how to interpret the value of *r*.
- talk about some of the properties of *r*.

Correlation coefficient (r)

(Definition)

The correlation coefficient (*r*) is a numeric measure that measures the *strength* and *direction* of a *linear* relationship between two quantitative variables.

Calculation: *r* is calculated using the following formula: 

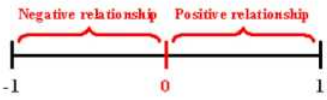
where n is the sample size; x is a data value for the explanatory variable;  is the mean of the x-values;  is the standard deviation of the x-values; similarly, for the terms involving y. To calculate r, the term  is calculated for each individual. These terms are added together, then the sum is divided by (n–1).

However, the calculation of *r* is not the focus of this course. We use a statistics package to calculate the correlation coefficient for us, and the emphasis of this course is on the *interpretation* of *r*'s value.

Interpretation

Once we obtain the value of *r*, its interpretation with respect to the strength of linear relationships is quite simple, as this walkthrough illustrates:

[redacted]



[redacted]

A YouTube element has been excluded from this version of the text. You can view it online here: pb.libretexts.org/cis/?p=154

Use the simulation below to investigate how the value of  relates to the direction and strength of the relationship between the two variables in the scatterplot.

In the simulation, use the slider bar at the top of the simulation to change the value of the correlation coefficient (*r*) between −1 and 1. Observe the effect on the scatterplot. Click on the "Switch Sign" button to jump between positive and negative relationships of

the same strength.

Click here to open this simulation in its own window.

A link to an interactive elements can be found at the bottom of this page.

Try It

https://assessments.lumenlearning.co...sessments/3477

https://assessments.lumenlearning.co...sessments/3478

https://assessments.lumenlearning.co...sessments/3479

https://assessments.lumenlearning.co...sessments/3480

https://assessments.lumenlearning.co...sessments/3481

https://assessments.lumenlearning.co...sessments/3482

Contributors and Attributions

3.22: Linear Relationships (3 of 4)

Learning Objectives

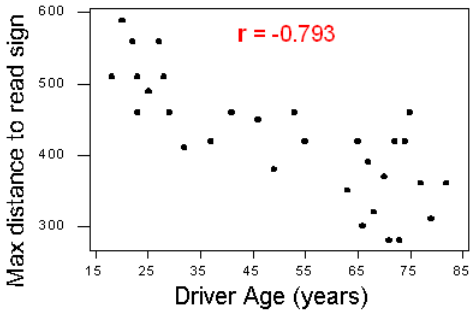
- Use a correlation coefficient to describe the direction and strength of a linear relationship. Recognize its limitations as a measure of the relationship between two quantitative variables.

Now we interpret the value of r in the context of some familiar examples.

Example

Highway Sign

In a previous example, we looked at this scatterplot to investigate the relationship between the age of a driver and the maximum distance at which the driver can read a highway sign. Because the form of the relationship is linear, we can use the correlation coefficient as a measure of direction and strength of the linear relationship.



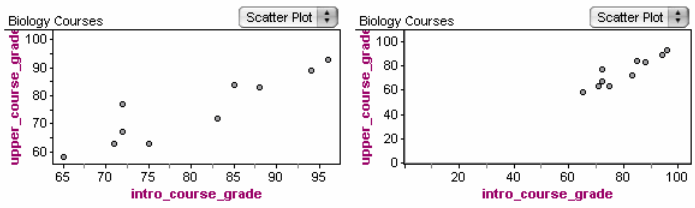
The r-value is −0.793. The r-value is negative ($r < 0$), which means that the linear relationship has a negative direction. We can see this in the scatterplot. Because r is somewhat close to −1, the relationship is moderately strong.

In the context of the data, the negative correlation confirms that the maximum reading distance decreases with age. Because r indicates a moderately strong linear relationship, we expect that drivers of similar age will have some (but not a lot) of variability in their maximum reading distance.

Example

Biology Courses

A biology department is interested in tracking the progress of its students from entry until graduation. As part of the study, the department tabulates the performance of 10 students in an introductory course and later in an upper-level course required for graduation. What is the relationship between the students' course grades in the two courses? Here are two scatterplots of the *same* data.



Both scatterplots show a relationship that is positive in direction and linear in form. The strength appears different in the two scatterplots because of the difference in scales. This illustrates why we support our visual assessment of strength with a measurement of strength. We can use the correlation coefficient as a measure of the strength of the linear relationship. The correlation coefficient is $r = 0.91$, which is close to 1. The correlation coefficient confirms that the linear relationship is very strong.

Comment

Note that in both examples, we supplemented the scatterplot with the correlation (r). Now that we have the correlation, why do we still need to look at a scatterplot when examining the relationship between two quantitative variables?

The correlation coefficient can be interpreted *only* as the *measure of the strength of a linear relationship*, so we need the scatterplot to verify that the relationship indeed looks linear. This point and its importance will be clearer after we examine a few properties of r.

Contributors and Attributions

3.23: Linear Relationships (4 of 4)

A link to an interactive elements can be found at the bottom of this page.

To see how an outlier affects the correlation, do the following:

1. Fill the scatterplot with a hypothetical positive linear relationship between X and Y (by clicking on the graph about a dozen times starting at the lower left and going up diagonally to the top right). Pay attention to the correlation coefficient calculated at the top left of the simulation. (Clicking on the garbage can lets you start over.)
2. Once you are satisfied with your hypothetical data, create an outlier by clicking on one of the data points in the upper right of the graph and dragging it down along the right side of the graph. Again, pay attention to what happens to the value of the correlation.

What did this activity illustrate? This activity illustrates that the correlation decreases when the outlier deviates from the pattern of the relationship. By dragging a data point from the upper right to the lower right, you created an outlier that does not fit the positive association in the rest of the data. This decreases the strength of the linear relationship and causes a decrease in r .

In the next activity, you will see how the correlation increases when the outlier is consistent with the direction of the linear relationship.

Let's Summarize

- A special case of the relationship between two quantitative variables is the **linear** relationship in which a straight line simply and adequately summarizes the relationship.
- When the scatterplot displays a linear relationship, we supplement it with the correlation coefficient (r), which measures the *strength* and *direction* of a linear relationship between two quantitative variables. The correlation ranges between −1 and 1. Values near −1 indicate a strong negative linear relationship, values near 0 indicate a weak linear relationship, and values near 1 indicate a strong positive linear relationship.
- The correlation is an appropriate numerical measure only for linear relationships and is sensitive to outliers. Therefore, the correlation should be used only as a supplement to a scatterplot (after we look at the data).

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.24: Introduction to Association vs Causation

What you'll learn to do: Distinguish between association and causation. Identify lurking variables that may explain an observed relationship.

Just because two variables are associated does not mean that one variable causes changes in the other! For example, swimsuit sales and beach toy sales are likely associated (as swimsuit sales go up, one might speculate that beach toy sales will also go up), but it's not necessarily the case that swimsuit sales cause beach toy sales.

In order to establish evidence of causation, a statistical study with rigorous design considerations is needed and the study results should be repeatable. We briefly discuss design considerations and appropriate conclusions that may be drawn.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.25: Causation and Lurking Variables (1 of 2)

Learning Objectives

- Distinguish between association and causation. Identify lurking variables that may explain an observed relationship.

Introduction

A common mistake people make when describing the relationship between two quantitative variables is that they confuse *association* and *causation*. This mistake is so common that we devote this entire section to clarifying the difference.

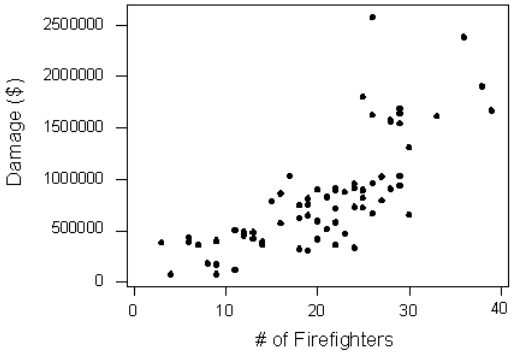
This confusion often occurs when there is a strong relationship between the two quantitative variables. In the case of a linear relationship, people mistakenly interpret an *r*-value that is close to 1 or -1 as evidence that the explanatory variable *causes* changes in the response variable. In this case, the *correct interpretation* is that there is a **statistical relationship** between the variables, not a causal link. In other words, the explanatory variable and the response variable vary together in a predictable way. There is an **association** between the variables. But this *should not* be interpreted as a cause-and-effect relationship.

Let's look at an example.

Example

Fire Damage

The scatterplot below shows the relationship between the number of firefighters sent to fires (x) and the amount of damage caused by fires (y) in a certain city.

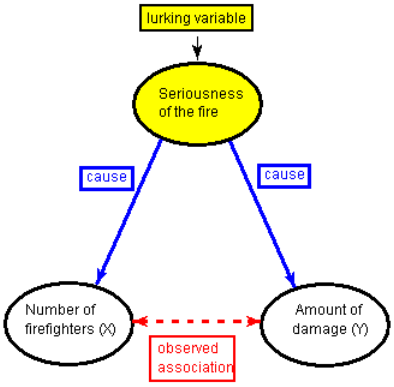


The scatterplot shows a positive association with a somewhat strong curvilinear form. An increase in the number of firefighters is associated with an increase in the damage done by the fire.

Can we conclude that the increase in firefighters causes the increase in damage? Of course not.

A third variable is at play in the background – the seriousness of the fire – and is responsible for the observed relationship. More serious fires require more firefighters and also result in more damage.

The following figure will help you visualize this situation:



The seriousness of the fire is a **lurking variable**. A lurking variable is a variable that is not measured in the study. It is a third variable that is neither the explanatory nor the response variable, but it affects your interpretation of the relationship between the explanatory and response variables.

In our example, the lurking variable has an effect on both the explanatory and the response variables. This common effect creates the observed association between the explanatory and response variables even though there is no cause-and-effect link between them.

Try It

https://assessments.lumenlearning.co...sessments/3851

Try It

https://assessments.lumenlearning.co...sessments/3852

Contributors and Attributions

3.26: Causation and Lurking Variables (2 of 2)

Learning Objectives

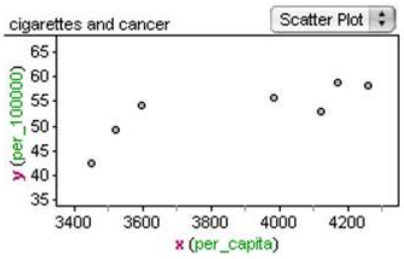
- Distinguish between association and causation. Identify lurking variables that may explain an observed relationship.

In the next example, we investigate a subtle point about the confusion between association and causation. In this example, a cause-and-effect connection is logical but not justified by an observed association in a single study.

Example

Smoking and Lung Cancer

In this data, x = cigarette consumption per capita in the United States, and y = lung cancers per 100,000. To investigate the connection between cigarette consumption and lung cancers, the data is offset by 30 years because cancer takes time to develop. For example, cigarette consumption in 1945 is paired with cancer rates for 1975.



In the scatterplot, we see a fairly strong positive correlation.

Can we conclude from this data that cigarette smoking causes lung cancer? The answer is no.

The data comes from an observational study. Recall from our previous discussions in Module 1 that we can draw cause-and-effect conclusions only from randomized comparative experiments. From this study, we can say that cigarette smoking is **associated** with lung cancer. We can also say that cigarette smoking **correlates** with lung cancer. We *cannot* say that cigarette smoking **causes** lung cancer.

Yet the National Cancer Institute's website states that "cigarette smoking causes many types of cancer, including cancers of the lung" (National Cancer Institute).

How can this be? Did the National Cancer Institute conduct a randomized comparative experiment to establish this cause-and-effect relationship? Of course not. We cannot randomly assign people to smoke or not smoke. All of the studies linking smoking with cancer are observational studies. Alone, each study can show only an association.

So is it possible to draw a causal link between cigarette consumption and cancer rates? The answer is yes, well sort of. In practice, researchers use criteria such as the following to provide evidence of a causal connection from observational studies:

- There is a reasonable explanation for how one variable might cause the other.
- The association is seen in repeated studies under varying conditions.
- The effects of potential lurking variables are ruled out when we look across studies.

The point of the previous example is again that association does not imply causation. But researchers can use an *observed association as the first step in building a case for causation.*

This point is subtle but important. When experiments cannot be conducted, it can be difficult and controversial to explain an observed association between two variables. Many of the current disputes involving data and statistics involve questions of causation that we cannot investigate through an experiment. Does the death penalty reduce violent crime? Does cell phone use cause brain tumors? Does pollution cause global warming? All of these questions imply a cause-and-effect relationship in situations that are complex and involve many interacting variables. In these situations, a single observational study cannot establish a causal link between two variables. But researchers can use the observed association as a first step in building a case for causation.

Try It

https://assessments.lumenlearning.co...sessments/3853

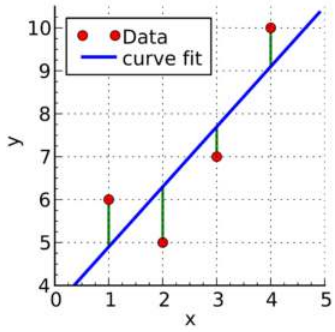
Let's Summarize

- The relationship between two quantitative variables is visually displayed using the *scatterplot*, where each point represents an individual. We always plot the explanatory variable on the horizontal axis and the response variable on the vertical axis.
- When we explore a relationship using the scatterplot, we should describe the *overall pattern* of the relationship and any *deviations* from that pattern. To describe the overall pattern, consider the *direction*, *form*, and *strength* of the relationship.
- Adding labels to the scatterplot that indicate different groups or categories within the data might help us gain more insight about the relationship we are exploring.
- A special case of the relationship between two quantitative variables is the *linear* relationship. In this case, a straight line simply and adequately summarizes the relationship.
- When the scatterplot displays a linear relationship, we supplement it with the *correlation coefficient (r)*, which measures the *strength* and the *direction* of a linear relationship between two quantitative variables. The correlation ranges between -1 and 1. Values near -1 indicate a strong negative linear relationship. Values near 0 can indicate a weak or no linear relationship. Values near 1 indicate a strong positive linear relationship. Remember, we use the correlation coefficient only *after* we have looked at the data and observed that there is a linear relationship. If you have no information about what the data actually looks like, then you should not use the correlation coefficient in your analysis.
- The correlation is an appropriate numerical measure only for linear relationships, and it is sensitive to outliers. Therefore, the correlation should be used only as a supplement to a scatterplot (after we look at the data).
- A *lurking variable* is a variable that is not measured in the study. It is a third variable that is neither the explanatory nor the response variable, but it affects your interpretation of the relationship between the explanatory and response variable.
- *Association does not imply causation.* Do not interpret a high correlation between explanatory and response variables as a cause-and-effect relationship.
- An observational study alone cannot establish a causal connection between explanatory and response variables. To establish a cause-and-effect relationship, researchers must conduct a comparative randomized experiment. In reality, it is often impossible to conduct an experiment. So observational studies that show an association between two variables can be used as a first step in building a case for causation.

Contributors and Attributions

3.27: Introduction to Linear Regression

What you'll learn to do: For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.



In this section, we present steps for finding the simple linear regression formula given a set of data. This formula is derived to find the line that has the smallest total squared error from the line to the observed data. In addition, we interpret the constants in a real-world context and explore the ways in which we can use the linear regression model to form predictions or good "guesses" for new values.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.28: Linear Regression (1 of 4)

Learning Objectives

- For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

So far we have used a scatterplot to describe the relationship between two quantitative variables. We described the pattern in the data by describing the direction, form, and strength of the relationship. We then focused on linear relationships. When the relationship is linear, we used correlation (r) as a measure of the direction and strength of the linear relationship.

Our focus on linear relationships continues here. We will

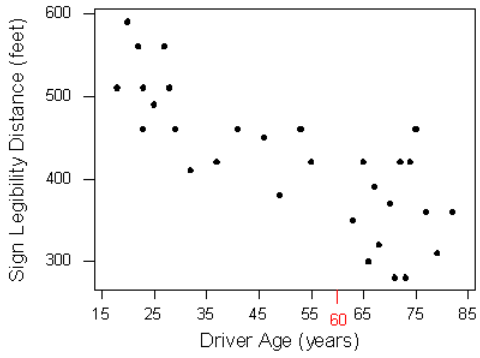
- use lines to make predictions.
- identify situations in which predictions can be misleading.
- develop a measurement for identifying the best line to summarize the data.
- use technology to find the best line.
- interpret the parts of the equation of a line to make our summary of the data more precise.

Making Predictions

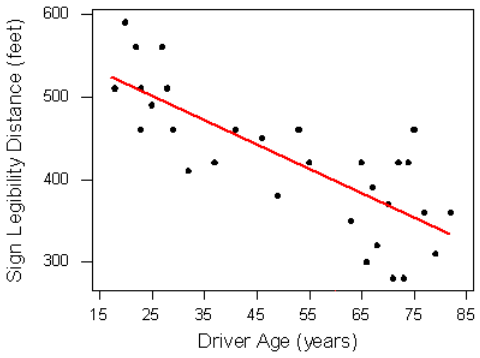
Earlier, we examined the linear relationship between the age of a driver and the maximum distance at which the driver can read a highway sign. Suppose we want to predict the maximum distance that a 60-year-old driver can read a highway sign. In the original data set, we do not have a 60-year-old driver.

How could we make a prediction using the linear pattern in the data?

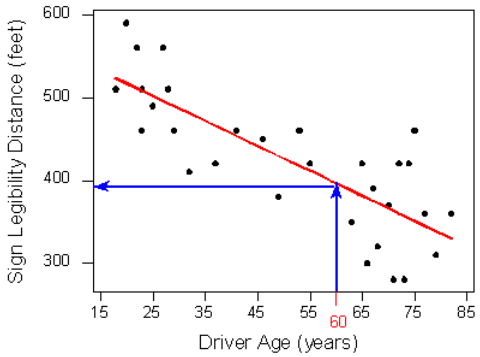
Here again is the scatterplot of driver ages and maximum reading distances . (Note: Sign Legibility Distance = Max distance to read sign.) We marked 60 on the x-axis.



Of course, different 60-year-olds will have different maximum reading distances . We expect variability among individuals. But here our goal is to make a single prediction that follows the general pattern in the data. Our first step is to model the pattern in the data with a line. In the scatterplot, you see a red line that follows the pattern in the data.



To use this line to make a prediction, we find the point on the line with an *x*-value of 60. Simply trace from 60 directly up to the line. We use the *y*-value of this point as the predicted maximum reading distance for a 60-year-old. Trace from this point across to the y-axis.



We predict that 60-year-old drivers can see the sign from a maximum distance of just under 400 feet.

We can also use the equation for the line to make a prediction. The equation for the red line is

Predicted distance = 576 − 3 * Age

To predict the maximum distance for a 60-year-old, substitute Age = 60 into the equation.

Predicted distance = 576 − 3 * (60) = 396 feet

Shortly, we develop a measurement for identifying the best line to summarize the data. We then use technology to find the equation of this line. Later, in "Assessing the Fit of a Line," we develop a method to measure the accuracy of the predictions from this "best" line. For now, just focus on how to use the line to make predictions.

Try It

https://assessments.lumenlearning.co...sessments/3489

https://assessments.lumenlearning.co...sessments/3490

Before we leave the idea of prediction, we end with the following cautionary note:

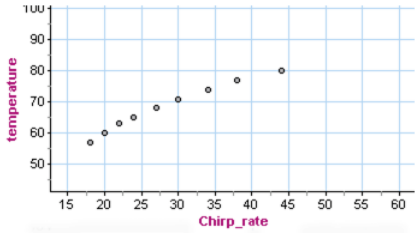
Avoid making predictions outside the range of the data.

Prediction for values of the explanatory variable that fall outside the range of the data is called **extrapolation**. These predictions are unreliable because we do not know if the pattern observed in the data continues outside the range of the data. Here is an example.

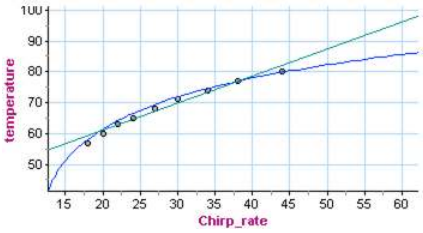
Example

Cricket Thermometers

Crickets chirp at a faster rate when the weather is warm. The scatterplot shows data presented in a 1995 issue of *Outside* magazine. Chirp rate is the number of chirps in 13 seconds. The temperature is in degrees Fahrenheit.



There is a strong relationship between chirp rate and temperature when the chirp rate is between about 18 and 45. What form does the data have? This is harder to determine. A line appears to summarize the data well, but we also see a curvilinear form, particularly when we pay attention to the first and last data points.



Both the curve and line are good summaries of the data. Both give similar predictions for temperature when the chirp rate is within the range of the data (between 18 and 45). But outside this range, the curve and the line give very different predictions. For example, if the crickets are chirping at a rate of 60, the line predicts a temperature just above 95°F. The curve predicts a much lower temperature of about 85°F.

Which is a better prediction? We do not know which is better because we do not know if the form is linear or curvilinear outside the range of the data.

If we use our model (the line or the curve) to make predictions outside the range of the data, this is an example of extrapolation. We see in this example that extrapolation can give unreliable predictions.

Try It

https://assessments.lumenlearning.co...sessments/3491

Try It

https://assessments.lumenlearning.co...sessments/3861

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.29: Linear Regression (2 of 4)

Learning Objectives

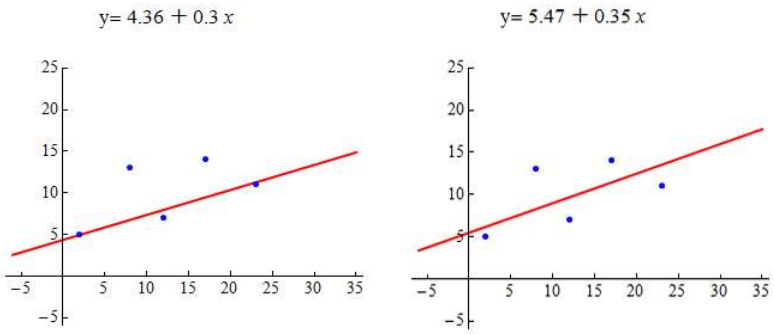
- For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

We continue our discussion of linear relationships with a focus on how to find the best line to summarize a linear pattern in data. Specifically, we do the following:

- Develop a measurement for identifying the best line to summarize the data.
- Use technology to find the best line.

Let's begin with a simple data set with only five data points.

Which line appears to be a better summary of the linear pattern in the data?



Let's make some observations about how these lines relate to the data points.

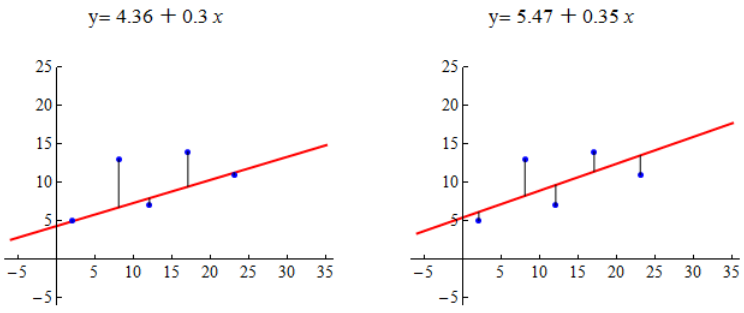
The line on the left passes through two of the five points. The point (12, 7) is very close to the line. The points (8, 13) and (17, 14) are relatively far from the line.

The line on the right does not pass through any of the points. It appears to pass through the middle of the distribution of the data. The points (8, 13) and (17, 14) are closer to this line than to the line on the left. But the other data points are farther from this line.

Which line is the best summary of the positive linear association we see in the data? Well, we may not agree on this, so we need a measurement of "best fit."

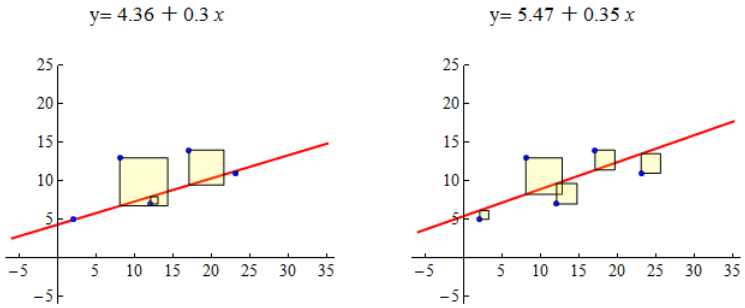
Here's the basic idea: The closer the line is to all of the data points, the better the line summarizes the pattern in the data. Notice when the line is close to the data points, it gives better predictions. A good prediction means the predicted *y*-value from the line is close to the actual *y*-value for the data point.

Here are the scatterplots again. For each data point, we drew a vertical line segment from the point to the summary line. The length of each vertical line segment is the amount that the predicted *y*-value deviates from the actual *y*-value for that data point. We think of this as the *error in the prediction*. We want to adjust the line until the overall error for all points together is as small as possible.



The most common measurement of overall error is the sum of the squares of the errors, or *SSE (sum of squared errors)*. The line with the smallest SSE is called the *least-squares regression line*. We call this line the "line of best fit."

Here are the scatterplots again. As before, each vertical line represents the error in a prediction. For each data point, the squared error is equal to the area of a yellow square. The least-squares regression line is the line with the smallest SSE, which means it has the smallest total yellow area.



Using the least-squares measurement, the line on the right is the better fit. It has a smaller sum of squared errors. When we compare the sum of the areas of the yellow squares, the line on the left has an SSE of 57.8. The line on the right has a smaller SSE of 43.9.

But is the line on the right the best fit? The answer is no. The line of best fit is the line that has the smallest sum of squared errors (SSE). For this data set, the line with the smallest SSE is $y = 6.72 + 0.26x$. The SSE is 41.79.

Now you try it with a new data set. Use the following simulation to adjust the line. See if you can find the least-squares regression line. (Try to find the line that makes the SSE as small as possible.)

Click here to open this simulation in its own window.

A link to an interactive elements can be found at the bottom of this page.

Try It

https://assessments.lumenlearning.co...sessments/3862

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

3.30: Linear Regression (3 of 4)

Learning Objectives

- For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

Let's quickly revisit the list of our data analysis tools for working with linear relationships:

- Use a scatterplot and *r* to describe direction and strength of the linear relationship.
- Find the equation of the least-squares regression line to summarize the relationship.
- Use the equation and the graph of the least-squares line to make predictions.
- Avoid extrapolation when making predictions.

Now we focus on the equation of a line in more detail. Our goal is to understand what the numbers in the equation tell us about the relationship between the explanatory variable and the response variable.

Here are some of the equations of lines that we have used in our discussion of linear relationships:

$$\text{Predicted distance} = 576 - 3 * \text{Age}$$

$$\text{Predicted height} = 39 + 2.7 * \text{forearm length}$$

$$\text{Predicted monthly car insurance premium} = 97 - 1.45 * \text{years of driving experience}$$

Notice that the form of the equations is the same. In general, each equation has the form

$$\text{Predicted } y = a + b * x$$

When we find the least-squares regression line, *a* and *b* are determined by the data. The values of *a* and *b* do not change, so we refer to them as **constants**.

In the equation of the line, the constant *a* is the prediction when $x = 0$. It is called **initial value**. In a graph of the line, *a* is the ***y*-intercept**.

In the equation of the line, the constant *b* is the rate of change, called the **slope**. In a graph of the least-squares line, *b* describes how the predictions change when *x* increases by one unit. More specifically, *b* describes the average change in the response variable when the explanatory variable increases by one unit.

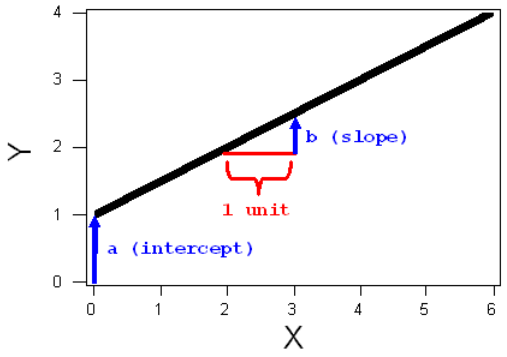
We can write the equation of the line to reflect the meaning of *a* and *b:*

$$\text{Predicted } y = a + b * x$$

$$\text{Predicted } y\text{-value} = (\text{initial value}) + (\text{rate of change}) * x$$

$$\text{Predicted } y\text{-value} = (y\text{-intercept}) + (\text{slope}) * x$$

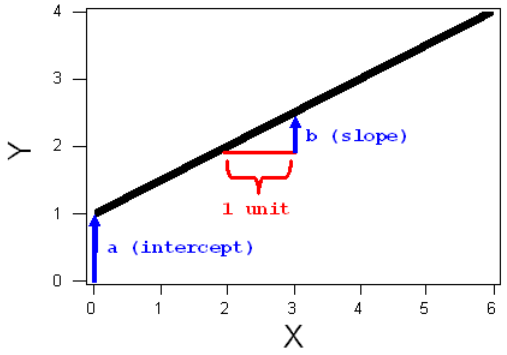
The constants *a* and *b* are shown in the graph of the line below.



Algebra review

The algebra of a line

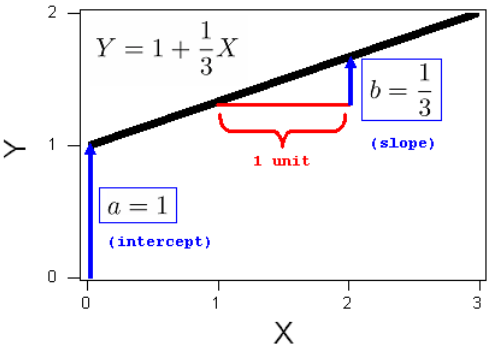
The general form for the equation of a line is Y = a + bX. The constants “a” and “b” can be either positive or negative. The constant “a” is the y-intercept where the line crosses the y-axis. The constant “b” is the slope. It describes the steepness of the line. In algebra we describe the slope as “rise over run”. The slope is the amount that Y increases (or decreases) for each 1-unit increase in X.



EXAMPLE

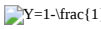
1

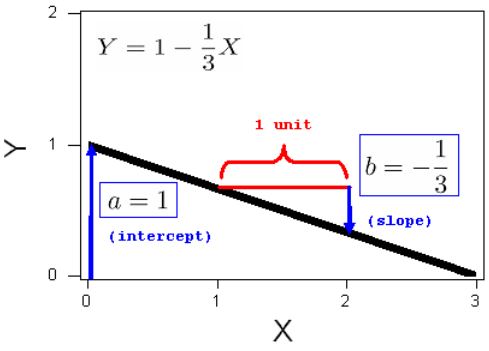
Consider the line  . The intercept is 1. The slope is 1/3, and the graph of this line is, therefore:



EXAMPLE

2

Consider the line  . The intercept is 1. The slope is -1/3, and the graph of this line is, therefore:



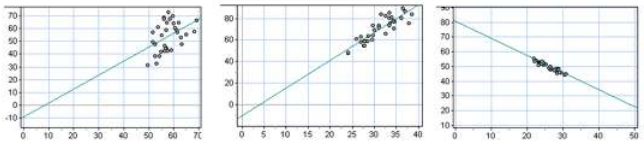
The simulation below allows you to see how changing the values of the slope and y-intercept changes the line. The slider on the left controls the y-intercept, a. The slider on the right controls the slope, b.

Use the simulation to draw the following lines:

$$Y = 3 + 0.67X$$
$$Y = 5 – X \text{ (which can also be written } Y = 5 – 1.0X)$$
$$Y = 2X \text{ (which can also be written } Y = 0 + 2X)$$
$$Y = 5 – 2X$$

A link to an interactive elements can be found at the bottom of this page.

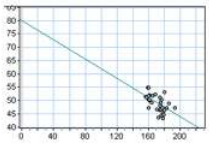
Use the following graphs in the next activity to investigate the equation of lines.



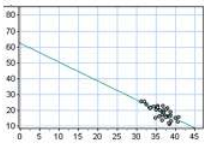
Graph A

Graph B

Graph C



Graph D



Graph E

Try It

https://assessments.lumenlearning.co...sessments/3483

https://assessments.lumenlearning.co...sessments/3492

Interpreting the Slope and Intercept

The constants in the equation of a line give us important information about the relationship between the predictions and x. In the next examples, we focus on how to interpret the meaning of the constants in the context of data.

Example

Highway Sign Visibility Data

Recall that from a data set of 30 drivers, we see a strong negative linear relationship between the age of a driver (x) and the maximum distance (in feet) at which a driver can read a highway sign. The least-squares regression line is

Predicted y-value = (starting value) + (rate of change)*x

$$\text{Predicted distance} = 576 - 3 * \text{Age}$$

$$\text{Predicted distance} = 576 + (-3 * \text{Age})$$

The value of b is −3. This means that a 1-year increase in age corresponds to a predicted 3-foot decrease in maximum distance at which a driver can read a sign. Another way to say this is that there is an average decrease of 3 feet in predicted sign visibility distance when we compare drivers of age x to drivers of age $x + 1$.

The 576 is the predicted value when $x = 0$. Obviously, it does not make sense to predict a maximum sign visibility distance for a driver who is 0 years old. This is an example of extrapolating outside the range of the data. But the starting value is an important part of the least-squares equation for predicting distances based on age.

The equation tells us that to predict the maximum visibility distance for a driver, start with a distance of 576 feet and subtract 3 feet for every year of the driver's age.

Example

Body Measurements

In the body measurement data collected from 21 female community college students, we found a strong positive correlation between forearm length and height. The least-squares regression line is

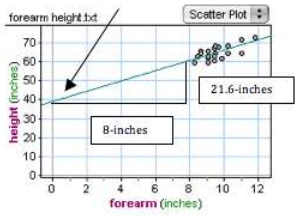
Predicted height = 39 + 2.7 * forearm length

The value of b is 2.7. This means that a 1-inch increase in forearm length corresponds to a predicted 2.7-inch increase in height. Another way to say this is that there is an average increase of 2.7-inches in predicted height when we compare women with forearm length of x to women with forearm length of $x + 1$.

The 39 is the predicted value when $x = 0$. Obviously, it does not make sense to predict the height of a woman with a 0-inch forearm length. This is another example of extrapolating outside the range of the data. But 39 inches is the starting value in the least-squares equation for predicting height based on forearm length.

The equation tells us that to predict the height of a woman, start with 39 inches and add 2.7 inches for every inch of forearm length.

In the graph below, we see the slope b represented by a triangle. An 8-inch increase in foreman length corresponds to a 21.6-inch increase in predicted height. $b = 21.6 / 8 = 2.7$. An arrow points to the starting value $a = 39$. This is the point with $x = 0$.



Try It

https://assessments.lumenlearning.co...sessments/3863

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

CHAPTER OVERVIEW

4: Relationships in Categorical Data with Intro to Probability

4.1: Why It Matters- Relationships in Categorical Data with Intro to Probability
4.2: Introduction to Two-Way Tables
4.3: Two-Way Tables (1 of 5)
4.4: Two-Way Tables (2 of 5)
4.5: Two-Way Tables (3 of 5)
4.6: Two-Way Tables (4 of 5)
4.7: Two-Way Tables (5 of 5)
4.8: Putting It Together- Relationships in Categorical Data with Intro to Probability

---

4.1: Why It Matters- Relationships in Categorical Data with Intro to Probability

Why understand the relationships within categorical data?

Before we begin *Relationships in Categorical Data with Intro to Probability*, it is helpful to consider how it relates to the work we have already done in previous modules.

At the start of *Summarizing Data Graphically and Numerically*, we stated the difference between quantitative and categorical variables:

- **Quantitative variables** have *numeric* values that can be averaged. A quantitative variable is frequently a measurement – for example, a person's height in inches.
- **Categorical variables** are variables that can have one of a limited number of values, or labels. Values that can be represented by categorical variables include, for example, a person's eye color, gender, or home state; a vehicle's body style (sedan, SUV, minivan, etc.); a dog's breed (bulldog, greyhound, beagle, etc.).

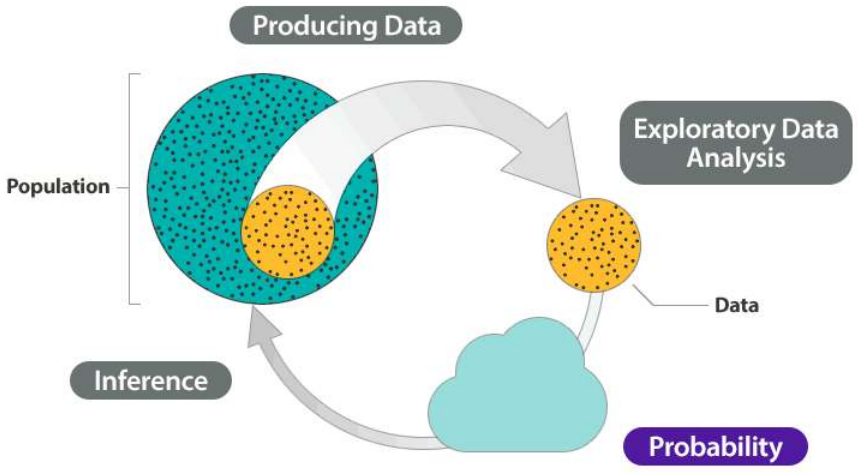
The remainder of *Summarizing Data Graphically and Numerically* focused on describing the overall pattern (shape, center, and spread) of the distribution of a quantitative variable.

In and *Examining Relationships: Quantitative Data* and *Nonlinear Models*, our goal was to identify and model the relationship between *two quantitative variables.*

Now, in this module, we turn our full attention back to categorical variables. Our objective is to study the relationship between two categorical variables. Just as in *Examining Relationships: Quantitative Data* and *Nonlinear Models*, we will be looking for patterns in the data.

As we organize and analyze data from two categorical variables, we make extensive use of **two-way tables**. Two-way tables for two categorical variables are in some ways like scatterplots for two quantitative variables: they give us a useful snapshot of all of the data organized in terms of the two variables of interest. This will be helpful in finding and comparing patterns. This part of *Relationships in Categorical Data with Intro to Probability* is exploratory data analysis in the Big Picture of Statistics.

A second important objective of this module is to introduce you to the concept of **probability**. Two-way tables give us a practical context for talking about probability. We also use two-way tables to help us visualize and solve real-world problems involving probability. This part of the module is part of probability in the Big Picture of Statistics.



Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

- **5.1: Why It Matters- Relationships in Categorical Data with Intro to Probability** by Lumen Learning is licensed CC BY 4.0.

4.2: Introduction to Two-Way Tables

What you'll learn to do: Analyze the relationship between two categorical variables using a two-way table.

Recall, categorical data is data that consists of labels (such as person's gender, an object's color, or location). Since categorical data does not return a measurement, it is often convenient to summarize study results with counts (for example, total number of females, or total number of males). In this section, we introduce two way tables and conditional percentages as a way to investigate possible relationships between two categorical variables.

	Arts-Sci	Bus-Econ	Info Tech	Health Science	Graphics Design	Culinary Arts	Row Totals
Female	4,660	435	494	421	105	83	6,198
Male	4,334	490	564	223	97	94	5,802
Column Totals	8,994	925	1,058	644	202	177	12,000

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

4.3: Two-Way Tables (1 of 5)

Learning Objectives

- Analyze the distribution of a categorical variable.
- Analyze the relationship between two categorical variables using a two-way table.

We begin our discussion by analyzing the distribution of a single categorical variable. Then we focus on analyzing the association between two categorical variables.

Example

Body Image

What is your perception of your own body? Do you feel that you are overweight, underweight, or about right? A random sample of 1,200 U.S. college students answered this question as part of a larger survey. The following table shows part of the responses:

Student	Body Image
student 25	overweight
student 26	about right
student 27	underweight
student 28	about right
student 29	about right

Here are the questions we investigate:

- What percentage of students in the sample fall into each category?
- How are students divided across the three body image categories?
- Is there a pattern in the responses?
- Which response is the most common?

It is difficult to answer these questions by looking at the raw data because the raw data is a long list of 1,200 responses. We cannot see patterns easily by looking at a list, so we summarize the distribution in a table.

Recall from *Summarizing Data Graphically and Numerically* that in a graph that summarizes the distribution of a *quantitative* variable, we can see

- the possible values of the variable.
- the number of individuals with each variable value or interval of values.

Here we use a table instead of a graph to summarize the distribution of a *categorical* variable. We create a table so we can see

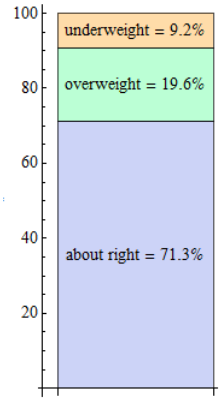
- the different values (categories) the variable takes.
- how many times each value occurs (count) and, more important, how often each value occurs (by converting the counts to proportions).

Here is the table for our example:

Category	Count	Proportion	Percentage
underweight	110	110/1,200 = 0.092	9.2%
overweight	235	235/1,200 = 0.196	19.6%
about right	855	855/1,200 = 0.713	71.3%

We can use a stacked bar chart to display the distribution of the body image variable. Note that this distribution is completely described by the three percentages 9.2%, 19.6%, and 71.3%, which correspond to the three categories of the body image variable:

“underweight,” “overweight,” and “about right.” The percentages add to 100% because all 1,200 individual responses fall into one of these three categories. (Note that the percentages actually add up to 99.9% because we rounded percentages to three decimal places.)



Now that we have summarized the distribution of values in the body image variable, let’s go back and interpret the results in the context of the questions we posed.

Try It

https://assessments.lumenlearning.co...sessments/3956

https://assessments.lumenlearning.co...sessments/3957

Example

Two-Way Table for Body Image and Gender

Once we’ve interpreted the results, another interesting question arises: *If we separate our sample by gender and compare the male and female responses, will we find a similar distribution across body image categories? Or is there a difference based on gender?*

Answering these questions requires us to examine the relationship between two categorical variables: gender and body image. We want to determine if gender explains the differences in body image responses. Therefore,

- the *explanatory* variable is gender, and
- the *response* variable is body image.

Here is part of the raw data for body image and gender of each student:

Student	Gender	Body Image
student 25	M	overweight
student 26	M	about right
student 27	F	underweight
student 28	F	about right
student 29	M	about right

Once again, the raw data is a long list of 1,200 responses. We need to organize the information in a table so we can more easily compare the results for females and males. To summarize the relationship between two categorical variables, we create a display called a **two-way table**.

Here is the two-way table for our example:

	About Right	Overweight	Underweight	Row Totals
Female	560	163	37	760

Male	295	72	73	440
Column Totals	855	235	110	1,200

Let's take a closer look at this table:

The table helps us to compare females to males because there is a row for each gender. The body image categories are the columns. As we move across a particular *row*, all of the individuals are of the *same gender*. And as we move down a particular *column*, all of the individuals have the *same body image*.

We also added a row at the bottom and a column at the right, which we call the **margins** of the table. The numbers in the margins are totals for each row or column.

In the following table, look at the numbers in the Female row and note that their sum, **560 + 163 + 37 = 760**, is displayed in the margin at the right labeled Row Totals. There are 760 females in the sample.

	About Right	**Overweight**	**Underweight**	**Row Totals**
Female	**560**	**163**	**37**	**760**
Male	295	72	73	440
Column Totals	855	235	110	1,200

Likewise, in the next table, look at the numbers in the Overweight column and note that their sum, **163 + 72 = 235**, is displayed in the margin at the bottom of the table labeled Column Totals. There are 235 students in the sample who answered "overweight" to the body image question.

	About Right	**Overweight**	**Underweight**	**Row Totals**
Female	560	**163**	37	760
Male	295	**72**	73	440
Column Totals	855	**235**	110	1,200

Where a row and column cross, we see the number of individuals who fit both descriptions: a particular gender and a particular body image. It may be helpful to think of the six inner cells as six rooms filled with the 1,200 students from the sample. For example, in one room are the 72 males who think of themselves as overweight. In another room, we have 37 females who think of themselves as underweight. (Maybe they should have a potluck and get to know each other.)

Try It

https://assessments.lumenlearning.co...sessments/3527

https://assessments.lumenlearning.co...sessments/3873

Try It

https://assessments.lumenlearning.co...sessments/3528

https://assessments.lumenlearning.co...sessments/3529

So far we have organized the raw data in a much more informative display – the two-way table. But we have not answered our primary question: Is body image related to gender?

Exploring the relationship between two categorical variables (in this case, body image and gender) amounts to *comparing the distributions of the response variable* (in this case, body image) *for different values of the explanatory variable* (in this case, male vs. female).

We do this in the next example.

Example

Is Body Image Related to Gender?

Here we have removed the column totals from the table because gender is the explanatory variable. We compare females with particular body image responses to males with the same response, so we need to know the total numbers of females and males. We no longer need to know the total number of students for each body image category.

Compare these distributions!

	about right	overweight	underweight	Row Totals
female	560	163	37	760
male	295	72	73	440

Note that there are more females than males, so when we compare females to males, it is misleading to compare raw counts in each body image category. For example, it is misleading to say, “Five-hundred sixty females responded ‘about right’ compared to only 295 males,” because the sample includes a lot more females than males. Instead, we compare the percentage of females who responded “about right” to the percentage of males who responded “about right”:

- Of the 760 females, 560 responded “about right”: 560 ÷ 760 = 0.737 = 73.7%
- Of the 440 males, 295 responded “about right”: 295 ÷ 440 = 0.67 = 67%

We can interpret percentages as “a number out of 100,” so by converting to percentages, we are reporting the results as though there are 100 females and 100 males. We can see that a higher percentage of females feel “about right” about their body weight.

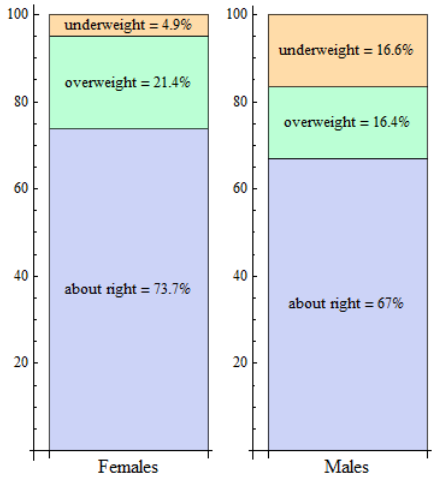
In general, we need to supplement our display, the two-way table, with numeric summaries that allow us to compare the distributions. Therefore, we always convert counts to percentages.

Note: It is important to identify the *explanatory* variable because we always use the totals for the explanatory variable to calculate the percentages.

In our example, we look at each gender separately and convert the counts to percentages within each gender. In the Female row, we divide each count by **760**, the total number of females. In the Male row, we divide each count by **440**, the total number of males. The resulting percentages are shown in the following table: green for females, black for males. We call these **conditional percentages**. The percentages in green are the distribution of body image based on the *condition that students are female*. The percentages in black are the distribution of body image based on the *condition that students are male*. Thus, our two sets of conditional percentages form two *conditional distributions* for body image.

	About Right	**Overweight**	**Underweight**	**Row Totals**
Female	560/760 = **73.7%**	163/760 = **21.4%**	37/760 =**4.9%**	760/760 = **100%**
Male	295/440 = **67%**	72/440 = **16.4%**	73/440 = **16.6%**	440/440 = **100%**

Here is a side-by-side display comparing the conditional body image distributions for females and males.



Now that we summarized the relationship between the categorical variables gender and body image, we use the next activity to interpret the results in the context of the questions we posed.

Try It

https://assessments.lumenlearning.co...sessments/3530

https://assessments.lumenlearning.co...sessments/3531

https://assessments.lumenlearning.co...sessments/3874

At the start of this example, we asked the following questions:

If we separate our sample by gender and compare the male and female responses, will we find a similar distribution across body image categories? Or is there a difference based on gender?

As a result of our analysis, we know that the conditional distributions for males and females for body image are not the same. And there is enough of a difference to believe that these two categorical variables are in fact related.

In the next activity, we practice investigating the relationship between two different categorical variables.

We investigate this question in the next activity: *Is there a relationship between smoking rates and college programs?* Researchers sent an online health behavior survey to 25,000 college students in 2009. The following table summarizes results based on 6,055 student responses. (C. J. Berg, C. M. Klatt, J. L. Thomas, J. S. Ahluwalia, and L. C. An, "The Relationship of Field of Study to Current Smoking Status among College Students," *College Student Journal* 43(3):744–754, 2009.)

	Smoked in Last 30 Days	Did Not Smoke in Last 30 Days	
Art, design, performing arts	149	336	485
Humanities	197	454	651
Communication, languages	233	389	622
Education	56	170	226
Health Sciences	227	717	944
Math, engineering, sciences	245	924	1,169
Social science, human services	306	593	899
Independent study	134	260	394
Undeclared	176	489	665
	1,723	4,332	6,055

Try It

https://assessments.lumenlearning.co...sessments/3532

https://assessments.lumenlearning.co...sessments/3533

https://assessments.lumenlearning.co...sessments/3534

https://assessments.lumenlearning.co...sessments/3535

In the next activity, we investigate whether health insurance coverage differs by geographic region. The U.S. government collects information on Americans who do not have health insurance. Here is the data:

Region	Uninsured	Insured	Row Totals
Northeast	6,782	47,043	53,825
Midwest	7,757	57,135	64,892
South	19,090	85,800	104,890
West	11,676	55,427	67,103
Column Totals	45,305	245,405	290,710

Let's Summarize

The relationship between two categorical variables is summarized using

- Data display: Two-way table, supplemented by
- Numeric summaries: Conditional percentages.

Conditional percentages are calculated separately for each value of the explanatory variable. When we try to understand the relationship between two categorical variables, we compare the distributions of the response variable for values of the explanatory variable. In particular, we look at how the pattern of conditional percentages differs between the values of the explanatory variable.

Contributors and Attributions

4.4: Two-Way Tables (2 of 5)

Learning Objectives

- Calculate marginal, joint, and conditional percentages and interpret them as probability estimates.

In the previous section, we used the information in a two-table to examine the relationship between two categorical variables. Our goal was to answer the big question: *Are the variables related?*

In this section, we continue to work with two-way tables, but we ask a different set of questions.

Example

Community College Enrollment

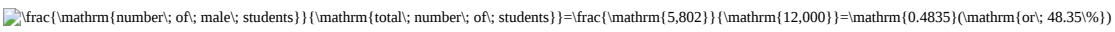
The following table summarizes the full-time enrollment at a community college located in a West Coast city. There are a total of 12,000 full-time students enrolled at the college. The two categorical variables here are *gender* and *program*. The programs include academic and vocational programs at the college. Assume that a student can enroll in only one program.

	Arts-Sci	**Bus-Econ**	**Info Tech**	**Health Science**	**Graphics Design**	**Culinary Arts**	**Row Totals**
Female	4,660	435	494	421	105	83	6,198
Male	4,334	490	564	223	97	94	5,802
Column Totals	8,994	925	1,058	644	202	177	12,000

Let's consider a few preliminary questions to get familiar with this new data set.

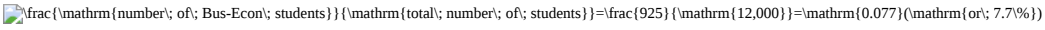
1. *What proportion of the total number of students are male students?*

Answer



2. *What proportion of the total number of students are Bus-Econ students?*

Answer



Note that to calculate this proportion, we used two numbers in the margin that relate to just one of the categorical variables (program). This calculation is therefore called a **marginal proportion**.

Note: This proportion does not help us determine if gender is related to program because it involves only one of the variables.

Now consider the following question:

> *If we choose one student at random from among all 12,000 students at the college, how likely is it that this student will be in the Bus-Econ program?*

From our previous calculation, we know that only about 8% (7.7%) of the students at the college are in the Bus-Econ program. That's a fairly low number, so it is not very likely that our random student will be a Bus-Econ student.

One way to state our conclusion is to say:

> There is about an **8% chance** of picking a Bus-Econ major.

This means that if we selected 100 students at random, we would expect on average that 8 of them would be in the Bus-Econ program.

Here is another way to state this conclusion:

> There is about an **0.08 probability** of picking a Bus-Econ major.

Because this probability is exactly the same as the marginal proportion we calculated earlier, we call it a **marginal probability**.

Note: *P* for Probability

It is customary to use the capital letter *P* to stand for probability. So instead of writing "The probability that a student is in Bus-Econ program equals 0.08," we can write *P*(student is in Bus-Econ) = 0.08.

The following table is used for the next Try It and Did I Get This? activities.

	Arts-Sci	Bus-Econ	Info Tech	Health Science	Graphics Design	Culinary Arts	Row Totals
Female	4,660	435	494	421	105	83	6,198
Male	4,334	490	564	223	97	94	5,802
Column Totals	8,994	925	1,058	644	202	177	12,000

Try It

https://assessments.lumenlearning.co...sessments/3536

https://assessments.lumenlearning.co...sessments/3537

Example

Conditional Probability

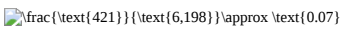
Here is the same community college enrollment data.

	Arts-Sci	Bus-Econ	Info Tech	Health Science	Graphics Design	Culinary Arts	Row Totals
Female	4,660	435	494	421	105	83	6,198
Male	4,334	490	564	223	97	94	5,802
Column Totals	8,994	925	1,058	644	202	177	12,000

Here is our first question:

> *If we select a female student at random, what is the probability that she is in the Health Sciences program?*

Answer Of the 6,198 female students at the college, **421** are enrolled in Health Sciences. (Find these numbers in the table.) The probability we are looking for is:



Therefore, the probability that a female student is in the Health Sciences program is approximately 0.07.

Focus on Language

We need to pause here and be very careful about the language we use in describing this situation.

Note that we *start* with a female student and *then* ask what is the probability that this female student is in the Health Sciences department.

In this case, our *starting point is that the student is a female*. This information sets the conditions for calculating the probability. Once the condition *(student is female)* is set, we focus on the female student population. In terms of the two-way table, it means that the only numbers we will be using are in the Female row: 421 and 6,198.

What Is a Conditional Probability?

The probability we calculated earlier is an example of a **conditional probability**. In general, a conditional probability is one that is based on a given condition. Here the *given condition* is that the student is female.

Here is the notation we use for a conditional probability:

- Original question: *If we select a female student at random, what is the probability that she is in the Health Sciences program?*
- Notation: *P*(student is in Health Sciences **given that** student is female).
- We also write this as *P*(Health Sciences **given** female).

An even shorter way of writing this is to use a vertical bar **|** in place of *given:P*(Health Sciences | female).

The following table is used for the next Try It and Did I Get This? activities.

	Arts-Sci	Bus-Econ	Info Tech	Health Science	Graphics Design	Culinary Arts	Row Totals
Female	4,660	435	494	421	105	83	6,198
Male	4,334	490	564	223	97	94	5,802
Column Totals	8,994	925	1,058	644	202	177	12,000

Try It

https://assessments.lumenlearning.co...sessments/3538

https://assessments.lumenlearning.co...sessments/3539

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

4.5: Two-Way Tables (3 of 5)

Learning Objectives

- Calculate marginal, joint, and conditional percentages and interpret them as probability estimates.

At this point, we know how to determine *marginal probabilities*, such as the probability that a randomly selected student is female: *P*(female).

And we know how to calculate *conditional probabilities*, such as the probability that a randomly selected female student is in the Health Science program: *P*(Health Science | female)

But we do not know how to calculate **joint probabilities**, such as the probability that a randomly selected student is both a female *and* in the Health Sciences program.

We write this joint probability as *P*(female and Health Sciences).

The following example illustrates how to calculate a joint probability.

Example

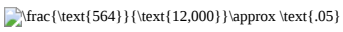
Joint Probability

Question: *If we select a student at random, what is the probability that the student is both a male **and** in the Info Tech program?*

Answer This question involves male students who are in the Info Tech program, but it is NOT a conditional probability. We are picking a student at random from the *entire population of 12,000 students*, so there is no condition. Our shorthand notation for this probability is:

P(male **and** Info Tech)

Since 564 of the 12,000 students enrolled at the college are both male and in the Info Tech program (see table), the probability *P*(male and Info Tech) is:



	Arts-Sci	Bus-Econ	Info Tech	Health Science	Graphics Design	Culinary Arts	Row Totals
Female	4,660	435	494	421	105	83	6,198
Male	4,334	490	**564**	223	97	94	5,802
Column Totals	8,994	925	1,058	644	202	177	**12,000**

We call this calculation a **joint probability**. Note that when we calculate a joint probability, we divide the count from an inner cell of the table by the overall total count in the lower right corner.

The following table is used for the next Try It and Did I Get This? activities.

	Arts-Sci	Bus-Econ	Info Tech	Health Science	Graphics Design	Culinary Arts	Row Totals
Female	4,660	435	494	421	105	83	6,198
Male	4,334	490	564	223	97	94	5,802
Column Totals	8,994	925	1,058	644	202	177	12,000

Try It

https://assessments.lumenlearning.co...sessments/3540

https://assessments.lumenlearning.co...sessments/3541

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

4.6: Two-Way Tables (4 of 5)

Learning Objectives

- Analyze and compare risks using conditional probabilities.

When we calculate the probability of a **negative outcome** like a heart attack, we often refer to the probability as a **risk**. For example, we talk about the *probability* of winning the lottery but the *risk* of getting struck by lightning. Whenever you see the word *risk*, keep in mind it's just another word for *probability*.

Example

Risk and the Physicians' Health Study

Researchers in the Physicians' Health Study (1989) designed a randomized clinical trial to determine whether aspirin reduces the risk of heart attack. Researchers randomly assigned a large sample of healthy male physicians (22,071) to one of two groups. One group took a low dose of aspirin (325 mg every other day). The other group took a placebo. This was a double-blind experiment. Here are the final results.

	Heart Attack	**No Heart Attack**	**Row Totals**
Aspirin	139	10,898	11,037
Placebo	239	10,795	11,034
Column Totals	378	21,693	22,071

Note that the categorical variables in this case are

- *Explanatory variable:* Treatment (aspirin or placebo)
- *Response variable:* Medical outcome (heart attack or no heart attack)

Question: *Does aspirin lower the risk of having a heart attack?*

To answer this question, we compare two conditional probabilities:

- The probability of a heart attack given that aspirin was taken every other day.
- The probability of a heart attack given that a placebo was taken every other day.

From the table we have

- *P*(heart attack | aspirin) = 139 / 11,037 = 0.013
- *P*(heart attack | placebo) = 239 / 11,034 = 0.022

The result shows that taking aspirin reduced the risk from 0.022 to 0.013.

We often compare two risks by calculating the **percentage change**. We calculate the difference (how much the risk changed) and divide by the risk for the placebo group.

Here is the calculation:

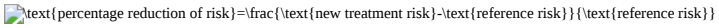
Therefore, we conclude that taking aspirin results in a 41% reduction in risk.

As reported in the *New England Journal of Medicine*, "This trial of aspirin for the primary prevention of cardiovascular disease demonstrates a conclusive reduction in the risk of myocardial infarction (heart attack)." *(Source: "Final Report on the Aspirin Component of the Ongoing Physicians' Health Study," New England Journal of Medicine 321(3):129–35, 1989.)*

Comment

In the preceding example, we compared the difference in risk (how much the risk changed) to the risk for the placebo (nontreatment) group:

In general, we are interested in determining how much a new treatment reduces the risk compared to a **reference** risk. The reference may be nontreatment (e.g., use of a placebo), or it could be an existing treatment that we hope to improve on. So we have:



The following table is used for the next Try It activity.

	Nonfatal	**Fatal**	**Row Totals**
Seat Belt	412,368	510	412,878
No Seat Belt	162,527	1,601	164,128
Column Totals	574,895	2,111	577,006

Try It

https://assessments.lumenlearning.co...sessments/3542

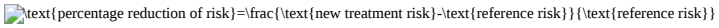
https://assessments.lumenlearning.co...sessments/3543

https://assessments.lumenlearning.co...sessments/3544

Let's summarize our work with probability. We defined three kinds of probabilities related to a two-way table.

- A *marginal probability* is the probability of a categorical variable taking on a particular value *without regard to the other categorical variable*. For example, *P*(Health Sciences) is the probability that a student is enrolled in the Health Sciences program. In calculating the probability, we use overall student data contained in the margins of the table. We do not take into account the other categorical variable: gender.
- A *conditional probability* is the probability of a categorical variable taking on a particular value *given the condition that the other categorical variable has some particular value*. For example, *P*(Health Sciences given female) is the probability that a student is enrolled in Health Sciences given that we know the student is female. In calculating the probability, we use only a subset of the data. The subset used is determined by the given condition: if our condition relates to female students, then we consider only the information in the table pertaining to females.
- A *joint probability* is the probability that the *two categorical variables each take on a specific value*. For example: *P*(male and Info Tech) is the probability that a student is both a male and in the Info Tech program. In calculating this probability, we divide the count in one inner cell of the table by the overall total count (in the lower right corner).

When we calculate the probability of a negative outcome like a heart attack, we often refer to the probability as a *risk*. We compare risk by calculating the percentage change:



Contributors and Attributions

4.7: Two-Way Tables (5 of 5)

Learning Objectives

- Create a hypothetical two-way table to answer more complex probability questions.

In our previous work with probability, we computed probabilities using a two-way table of data from a large sample. Now we create a hypothetical two-way table to answer more complex probability questions.

Example

Will It Be a Boy or a Girl?

A pregnant woman often opts to have an ultrasound to predict the gender of her baby.

Assume the following facts are known:

- Fact 1: 48% of the babies born are female.
- Fact 2: The proportion of girls correctly identified is 9 out of 10.
- Fact 3: The proportion of boys correctly identified is 3 out of 4.

(Source: Keeler, Carolyn, and Steinhorst, Kirk. "New Approaches to Learning Probability in the First Statistics Course," Journal of Statistics Education 9(3):1–24, 2001.)

Here are the questions we want to answer:

Question 1: *If the examination predicts a girl, how likely is it that the baby will be a girl?*
Question 2: *If the examination predicts a boy, how likely is it that the baby will be a boy?*

Let's consider what the possibilities are.

- The ultrasound examination predicts a girl, and either (a) a girl is born or (b) a boy is born.
- The ultrasound exam predicts a boy, and either (a) a girl is born or (b) a boy is born.

Let's represent these four possible outcomes in a two-way table. On the left we have the categorical variable *prediction*, and on the top the categorical variable *gender of baby*.

	Girl	**Boy**	
Predict Girl			
Predict Boy			

Now we find ourselves in an interesting situation. A two-way table without data!

The key idea is to create a two-way table consistent with the stated facts, then use the table to answer our questions.

To get started, let's assume we have ultrasound predictions for 1,000 random babies. We could have picked any number here, but 1,000 will make our calculations easier to keep track of.

Starting with this number, we work backwards with our three facts to fill in this "hypothetical" table.

The first step is to put 1,000 as the overall total in the bottom right corner.

	Girl	**Boy**	**Row Totals**
Predict Girl			
Predict Boy			
Column Totals			**1,000**

Let's consider Fact 1: 48% of the babies born are female.

The bottom row gives the distribution of the categorical variable *gender of baby*. We can use this fact to compute the total number of girls and boys.

- 48% girls means that 0.48 (1,000) = 480 are girls.
- 52% are boys. (If 48% are girls, then 100% − 48% = 52% are boys.) So, 0.52(1,000) = 520 boys.

Fill these values into the bottom row of table.

- Note: These are marginal totals.
- You can check your work: These numbers should add to 1,000. If we add all the girls and boys together, we get the total number of babies.

	Girl	**Boy**	**Row Totals**
Predict Girl			
Predict Boy			
Column Totals	0.48(1,000) = **480**	0.52(1,000) = **520**	1,000

Now let's move on to Fact 2: The proportion of girls correctly identified is 9 out of 10.

- 9 out of 10 is 90% (9 ÷ 10 = 0.90 = 90%).
- 90% of the girls are correctly identified: 0.90(480) = 432.
- 10% of the girls are misidentified (predicted to be a boy): 0.10(480) = 48.

Fill these values into the table.

- You can check your work: These numbers should add to the total number of girls.
- (Girls who are correctly identified as girls) + (Girls who are misidentified as boys) = Total girls

	Girl	**Boy**	**Row Totals**
Predict Girl	0.90(480)= **432**		
Predict Boy	0.10(480) = **48**		
Column Totals	480	520	1,000

Finally, we use Fact 3: The proportion of boys correctly identified is 3 out of 4.

- 3 out of 4 is 75% (3 ÷ 4 = 0.75 = 75%).
- 75% of the boys are correctly identified: 0.75(520) = 390.
- 25% of the boys are misidentified (predicted to be a girl): 0.25(520) = 130.

Fill these values into the table.

- You can check your work: These numbers should add to the total number of boys.
- (Boys who are correctly identified as boys) + (Boys who are misidentified as girls) = Total boys

	Girl	**Boy**	**Row Totals**
Predict Girl	432	0.25(520) = **130**	
Predict Boy	48	0.75(520) = **390**	
Column Totals	480	520	1,000

Filling in the Row Totals, we now have a complete hypothetical two-way table based on our given information.

	Girl	**Boy**	**Row Totals**
Predict Girl	432	130	562

Predict Boy	48	390	438
Column Totals	480	520	1,000

We are now in a position to answer our two questions:

Question 1: *If the examination predicts a girl, how likely is it that the baby will be a girl?*

Answer: We are asked to find the probability of a girl **given that** the examination predicts a girl.

This is the conditional probability: *P*(girl | predict girl).

So our answer to Question 1 is *P*(girl | predict girl) = 432 / 562 = 0.769.

Question 2: *If the examination predicts a boy, how likely is it that the baby will be a boy?*

Answer: We are asked to find the probability of a boy **given that** the examination predicts a boy.

This is the conditional probability: *P*(boy | predict boy).

So our answer to Question 2 is *P*(boy | predict boy) = 390 / 438 = 0.890.

Conclusion: If an ultrasound examination predicts a girl, the prediction is correct about 77% of the time. In contrast, when the prediction is a boy, it is correct 89% of the time.

Comment

Are you surprised at the answers to these questions? Looking just at the three given facts, you might have intuitively expected a different result. This is exactly why a two-way table is so useful. It helps us organize the relevant information in a way that permits us to carry out a logical analysis. When it comes to probability, sometimes our intuition needs some help.

Use the following context for the next Try It activity.

A large company has instituted a mandatory employee drug screening program. Assume that the drug test used is known to be 99% accurate. That is, if an employee is a drug user, the test will come back positive (“drug detected”) 99% of the time. If an employee is a non-drug user, then the test will come back negative (“no drug detected”) 99% of the time. Assume that 2% of the employees of the company are drug users.

In constructing the hypothetical two-way table, it is convenient to start by assuming that the company has 10,000 employees (10,000 is a large enough number to ensure that all calculations result in whole numbers).

Try It

https://assessments.lumenlearning.co...sessments/3545

https://assessments.lumenlearning.co...sessments/3546

https://assessments.lumenlearning.co...sessments/3547

https://assessments.lumenlearning.co...sessments/3548

https://assessments.lumenlearning.co...sessments/3549

https://assessments.lumenlearning.co...sessments/3550

Contributors and Attributions

4.8: Putting It Together- Relationships in Categorical Data with Intro to Probability

Let's Summarize

To summarize the relationship between two categorical variables, use:

- A data display: A two-way table
- Numerical summaries: Conditional percentages

When we investigate the relationship between two categorical variables, we use the values of the explanatory variable to define the comparison groups. We then compare the distributions of the response variable for values of the explanatory variable. In particular, we look at how the pattern of conditional percentages differs between the values of the explanatory variable.

> For example, we investigated the relationship between body image and gender. We compared males to females. For each gender, we determined the percentage who felt their body weight was about right, overweight, or underweight. *P*(body image "about right" | male) is compared to *P*(body image "about right" | female).

Keys Ideas from Our Work with Probability

We defined three kinds of probabilities related to a two-way table:

- A **marginal probability** is the probability of a categorical variable taking on a particular value *without regard to the other categorical variable*. For example, *P*(Health Sciences) is the probability that a student is enrolled in the Health Sciences program. In calculating the probability, we use overall student data contained in the margins of the table. A marginal probability is a row or column total divided by the table total.
- A **conditional probability** is the probability of a categorical variable taking on a particular value *given the condition that the other categorical variable has some particular value*. For example, *P*(Health Sciences **given** female) means we look first at all females, then identify the female students who are Health Science students. In calculating the probability, we use only a subset of the data. The condition determines the subset of data we use. If our condition relates to female students, then we consider only the information in the table pertaining to females.
- A **joint probability** is the probability that the *two categorical variables each take on a specific value*. For example: *P*(male **and** Info Tech) is the probability that a student is both a male and in the Info Tech program. In calculating this probability, we divide the count from one inner cell of the table by the overall total count (in the lower right corner.)

When we calculate the probability of a **negative outcome**, we often refer to the probability as a **risk**. We compare risk by calculating the percentage change (divide difference in risks by risk in placebo group).

Finally, we created hypothetical two-way tables to compute complex probabilities, such as the probability of a positive drug test for someone who does not use drugs.

Contributors and Attributions

CHAPTER OVERVIEW

5: Basic Concepts of Probability

Suppose a polling organization questions 1,200 voters in order to estimate the proportion of all voters who favor a particular bond issue. We would expect the proportion of the 1,200 voters in the survey who are in favor to be close to the proportion of all voters who are in favor, but this need not be true. There is a degree of randomness associated with the survey result. If the survey result is highly likely to be close to the true proportion, then we have confidence in the survey result. If it is not particularly likely to be close to the population proportion, then we would perhaps not take the survey result too seriously. The likelihood that the survey proportion is close to the population proportion determines our confidence in the survey result. For that reason, we would like to be able to compute that likelihood. The task of computing it belongs to the realm of probability, which we study in this chapter.

5.1: Sample Spaces, Events, and Their Probabilities

5.2: Complements, Intersections, and Unions

5.3: Conditional Probability and Independent Events

5.E: Basic Concepts of Probability (Exercises)

5.1: Sample Spaces, Events, and Their Probabilities

Learning Objectives

- To learn the concept of the sample space associated with a random experiment.
- To learn the concept of an event associated with a random experiment.
- To learn the concept of the probability of an event.

Sample Spaces and Events

Rolling an ordinary six-sided die is a familiar example of a *random experiment*, an action for which all possible outcomes can be listed, but for which the actual outcome on any given trial of the experiment cannot be predicted with certainty. In such a situation we wish to assign to each outcome, such as rolling a two, a number, called the *probability* of the outcome, that indicates how likely it is that the outcome will occur. Similarly, we would like to assign a probability to any *event*, or collection of outcomes, such as rolling an even number, which indicates how likely it is that the event will occur if the experiment is performed. This section provides a framework for discussing probability problems, using the terms just mentioned.

Definition: random experiment

A *random experiment* is a mechanism that produces a definite outcome that cannot be predicted with certainty. The sample space associated with a random experiment is the set of all possible outcomes. An event is a subset of the sample space.

Definition: Element and Occurrence

An event E is said to occur on a particular trial of the experiment if the outcome observed is an element of the set E.

Example 5.1.1: Sample Space for a single coin

Construct a sample space for the experiment that consists of tossing a single coin.

Solution

The outcomes could be labeled h for heads and t for tails. Then the sample space is the set: $S = \{h, t\}$

Example 5.1.2: Sample Space for a single die

Construct a sample space for the experiment that consists of rolling a single die. Find the events that correspond to the phrases "an even number is rolled" and "a number greater than two is rolled."

Solution

The outcomes could be labeled according to the number of dots on the top face of the die. Then the sample space is the set $S = \{1, 2, 3, 4, 5, 6\}$

The outcomes that are even are $2, 4,$ and 6, so the event that corresponds to the phrase "an even number is rolled" is the set $\{2, 4, 6\}$, which it is natural to denote by the letter E. We write $E = \{2, 4, 6\}$.

Similarly the event that corresponds to the phrase "a number greater than two is rolled" is the set $T = \{3, 4, 5, 6\}$, which we have denoted T.

A graphical representation of a sample space and events is a *Venn diagram*, as shown in Figure 5.1.1. In general the sample space S is represented by a rectangle, outcomes by points within the rectangle, and events by ovals that enclose the outcomes that compose them.

(a): One Coin Toss (b): One Die Roll

Figure 5.1.1: Venn Diagrams for Two Sample Spaces

Example 5.1.3: Sample Spaces for two coines

A random experiment consists of tossing two coins.

a. Construct a sample space for the situation that the coins are indistinguishable, such as two brand new pennies.
b. Construct a sample space for the situation that the coins are distinguishable, such as one a penny and the other a nickel.

Solution

a. After the coins are tossed one sees either two heads, which could be labeled $2h$, two tails, which could be labeled $2t$, or coins that differ, which could be labeled d Thus a sample space is $S = \{2h, 2t, d\}$.
b. Since we can tell the coins apart, there are now two ways for the coins to differ: the penny heads and the nickel tails, or the penny tails and the nickel heads. We can label each outcome as a pair of letters, the first of which indicates how the penny landed and the second of which indicates how the nickel landed. A sample space is then $S' = \{hh, ht, th, tt\}$.

A device that can be helpful in identifying all possible outcomes of a random experiment, particularly one that can be viewed as proceeding in stages, is what is called a *tree diagram*. It is described in the following example.

Example 5.1.4: Tree diagram

Construct a sample space that describes all three-child families according to the genders of the children with respect to birth order.

Solution

Two of the outcomes are "two boys then a girl," which we might denote bbg, and "a girl then two boys," which we would denote gbb.

Clearly there are many outcomes, and when we try to list all of them it could be difficult to be sure that we have found them all unless we proceed systematically. The tree diagram shown in Figure 5.1.2, gives a systematic approach.

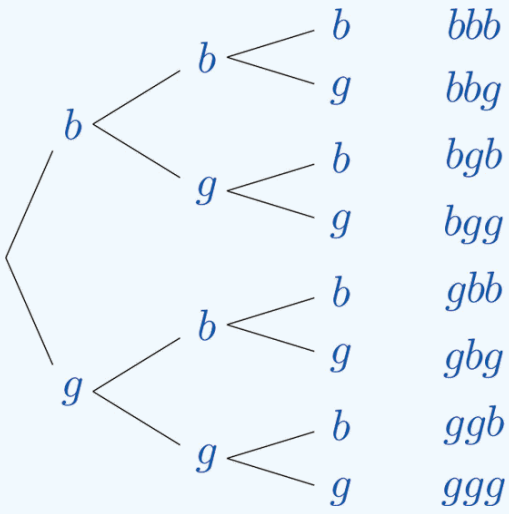


Figure 5.1.2: *Tree Diagram For Three-Child Families*

The diagram was constructed as follows. There are two possibilities for the first child, boy or girl, so we draw two line segments coming out of a starting point, one ending in a b for "boy" and the other ending in a g for "girl." For each of these two possibilities

for the first child there are two possibilities for the second child, “boy” or “girl,” so from each of the b and g we draw two line segments, one segment ending in a b and one in a g. For each of the four ending points now in the diagram there are two possibilities for the third child, so we repeat the process once more.

The line segments are called **branches** of the tree. The right ending point of each branch is called a **node**. The nodes on the extreme right are the **final nodes**; to each one there corresponds an outcome, as shown in the figure.

From the tree it is easy to read off the eight outcomes of the experiment, so the sample space is, reading from the top to the bottom of the final nodes in the tree,

$$S = \{bbb,\ bbg,\ bgb,\ bgg,\ gbb,\ gbg,\ ggb,\ ggg\}$$

Probability

Definition: probability

The probability of an outcome e in a sample space S is a number P between 1 and 0 that measures the likelihood that e will occur on a single trial of the corresponding random experiment. The value $P = 0$ corresponds to the outcome e being impossible and the value $P = 1$ corresponds to the outcome e being certain.

Definition: probability of an event

The *probability of an event* A is the sum of the probabilities of the individual outcomes of which it is composed. It is denoted $P(A)$.

The following formula expresses the content of the definition of the probability of an event:

If an event E is $E = \{e_1, e_2, \ldots, e_k\}$, then

$$P(E) = P(e_1) + P(e_2) + \ldots + P(e_k)$$

The following figure expresses the content of the definition of the probability of an event:

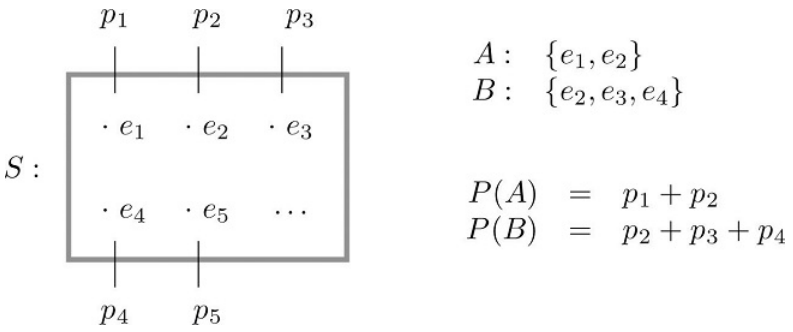


Figure 5.1.3: *Sample Spaces and Probability*

Since the whole sample space S is an event that is certain to occur, the sum of the probabilities of all the outcomes must be the number 1.

In ordinary language probabilities are frequently expressed as percentages. For example, we would say that there is a 70% chance of rain tomorrow, meaning that the probability of rain is 0.70. We will use this practice here, but in all the computational formulas that follow we will use the form 0.70 and not 70%.

Example 5.1.5

A coin is called “balanced” or “fair” if each side is equally likely to land up. Assign a probability to each outcome in the sample space for the experiment that consists of tossing a single fair coin.

Solution

With the outcomes labeled h for heads and t for tails, the sample space is the set

$$S = \{h, t\}$$

Since the outcomes have the same probabilities, which must add up to 1, each outcome is assigned probability $1/2$.

Example 5.1.6

A die is called "balanced" or "fair" if each side is equally likely to land on top. Assign a probability to each outcome in the sample space for the experiment that consists of tossing a single fair die. Find the probabilities of the events E: "an even number is rolled" and T: "a number greater than two is rolled."

Solution

With outcomes labeled according to the number of dots on the top face of the die, the sample space is the set

$$S = \{1, 2, 3, 4, 5, 6\}$$

Since there are six equally likely outcomes, which must add up to 1, each is assigned probability $1/6$.

Since $E = \{2, 4, 6\}$,

$$P(E) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Since $T = \{3, 4, 5, 6\}$,

$$P(T) = \frac{4}{6} = \frac{2}{3}$$

Example 5.1.7

Two fair coins are tossed. Find the probability that the coins match, i.e., either both land heads or both land tails.

Solution

In Example 5.1.3 we constructed the sample space $S = \{2h, 2t, d\}$ for the situation in which the coins are identical and the sample space $S' = \{hh, ht, th, tt\}$ for the situation in which the two coins can be told apart.

The theory of probability does not tell us how to assign probabilities to the outcomes, only what to do with them once they are assigned. Specifically, using sample space S, matching coins is the event $M = \{2h, 2t\}$ which has probability $P(2h) + P(2t)$. Using sample space S', matching coins is the event $M' = \{hh, tt\}$, which has probability $P(hh) + P(tt)$. In the physical world it should make no difference whether the coins are identical or not, and so we would like to assign probabilities to the outcomes so that the numbers $P(M)$ and $P(M')$ are the same and best match what we observe when actual physical experiments are performed with coins that seem to be fair. Actual experience suggests that the outcomes in S' are equally likely, so we assign to each probability $\frac{1}{4}$, and then...

$$P(M') = P(hh) + P(tt) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Similarly, from experience appropriate choices for the outcomes in S are:

$$P(2h) = \frac{1}{4}$$

$$P(2t) = \frac{1}{4}$$

$$P(d) = \frac{1}{2}$$

The previous three examples illustrate how probabilities can be computed simply by counting when the sample space consists of a finite number of equally likely outcomes. In some situations the individual outcomes of any sample space that represents the experiment are unavoidably unequally likely, in which case probabilities cannot be computed merely by counting, but the computational formula given in the definition of the probability of an event must be used.

Example 5.1.8

The breakdown of the student body in a local high school according to race and ethnicity is 51% white, 27% black, 11% Hispanic, 6% Asian, and 5% for all others. A student is randomly selected from this high school. (To select "randomly" means that every student has the same chance of being selected.) Find the probabilities of the following events:

a. B: the student is black,
b. M: the student is minority (that is, not white),
c. N: the student is not black.

Solution

The experiment is the action of randomly selecting a student from the student population of the high school. An obvious sample space is $S = \{w, b, h, a, o\}$. Since 51% of the students are white and all students have the same chance of being selected, $P(w) = 0.51$, and similarly for the other outcomes. This information is summarized in the following table:

$Outcome$	w	b	h	a	o
$Probability$	0.51	0.27	0.11	0.06	0.05

a. Since $B = \{b\}$, $P(B) = P(b) = 0.27$
b. Since $M = \{b, h, a, o\}$, $P(M) = P(b) + P(h) + P(a) + P(o) = 0.27 + 0.11 + 0.06 + 0.05 = 0.49$
c. Since $N = \{w, h, a, o\}$, $P(N) = P(w) + P(h) + P(a) + P(o) = 0.51 + 0.11 + 0.06 + 0.05 = 0.73$

Example 5.1.9

The student body in the high school considered in the last example may be broken down into ten categories as follows: 25% white male, 26% white female, 12% black male, 15% black female, 6% Hispanic male, 5% Hispanic female, 3% Asian male, 3% Asian female, 1% male of other minorities combined, and 4% female of other minorities combined. A student is randomly selected from this high school. Find the probabilities of the following events:

a. B: the student is black
b. MF: the student is a non-white female
c. FN: the student is female and is not black

Solution

Now the sample space is $S = \{wm, bm, hm, am, om, wf, bf, hf, af, of\}$ The information given in the example can be summarized in the following table, called a two-way contingency table:

Gender	Race / Ethnicity				
	White	Black	Hispanic	Asian	Others
Male	0.25	0.12	0.06	0.03	0.01
Female	0.26	0.15	0.05	0.03	0.04

a. Since $B = \{bm, bf\}$, $P(B) = P(bm) + P(bf) = 0.12 + 0.15 = 0.27$
b. Since $MF = \{bf, hf, af, of\}$, $P(M) = P(bf) + P(hf) + P(af) + P(of) = 0.15 + 0.05 + 0.03 + 0.04 = 0.27$
c. Since
$FN = \{wf, hf, af, of\}$, $P(FN) = P(wf) + P(hf) + P(af) + P(of) = 0.26 + 0.05 + 0.03 + 0.04 = 0.38$

Key Takeaway

- The sample space of a random experiment is the collection of all possible outcomes.
- An event associated with a random experiment is a subset of the sample space.
- The probability of any outcome is a number between 0 and 1. The probabilities of all the outcomes add up to 1.
- The probability of any event A is the sum of the probabilities of the outcomes in A.

5.2: Complements, Intersections, and Unions

Learning Objectives

- To learn how some events are naturally expressible in terms of other events.
- To learn how to use special formulas for the probability of an event that is expressed in terms of one or more other events.

Some events can be naturally expressed in terms of other, sometimes simpler, events.

Complements

Definition: Complement

The complement of an event A in a sample space S, denoted A^c, is the collection of all outcomes in S that are not elements of the set A. It corresponds to negating any description in words of the event A.

Example 5.2.1

Two events connected with the experiment of rolling a single die are E: ***"the number rolled is even"*** and T: ***"the number rolled is greater than two."*** Find the complement of each.

Solution

In the sample space $S = \{1, 2, 3, 4, 5, 6\}$ the corresponding sets of outcomes are $E = \{2, 4, 6\}$ and $T = \{3, 4, 5, 6\}$. The complements are $E^c = \{1, 3, 5\}$ and $T^c = \{1, 2\}$.

In words the complements are described by "the number rolled is not even" and "the number rolled is not greater than two." Of course easier descriptions would be "the number rolled is odd" and "the number rolled is less than three."

If there is a 60% chance of rain tomorrow, what is the probability of fair weather? The obvious answer, 40%, is an instance of the following general rule.

Definition: Probability Rule for Complements

The Probability Rule for Complements states that

$$P(A^c) = 1 - P(A)$$

This formula is particularly useful when finding the probability of an event directly is difficult.

Example 5.2.2

Find the probability that at least one heads will appear in five tosses of a fair coin.

Solution

Identify outcomes by lists of five hs and ts, such as $tthtt$ and $hhttt$. Although it is tedious to list them all, it is not difficult to count them. Think of using a tree diagram to do so. There are two choices for the first toss. For each of these there are two choices for the second toss, hence $2 \times 2 = 4$ outcomes for two tosses. For each of these four outcomes, there are two possibilities for the third toss, hence $4 \times 2 = 8$ outcomes for three tosses. Similarly, there are $8 \times 2 = 16$ outcomes for four tosses and finally $16 \times 2 = 32$ outcomes for five tosses.

Let O denote the event "at least one heads." There are many ways to obtain at least one heads, but only one way to fail to do so: all tails. Thus although it is difficult to list all the outcomes that form O, it is easy to write $O^c = \{ttttt\}$. Since there are 32 equally likely outcomes, each has probability $\frac{1}{32}$, so $P(O^c) = 1 / 32$, hence $P(O) = 1 - \frac{1}{32} \approx 0.97$ or about a 97% chance.

Intersection of Events

Definition: intersections

The intersection of events A and B, denoted $A \cap B$, is the collection of all outcomes that are elements of both of the sets A and B. It corresponds to combining descriptions of the two events using the word "and."

To say that the event $A \cap B$ occurred means that on a particular trial of the experiment both A and B occurred. A visual representation of the intersection of events A and B in a sample space S is given in Figure 5.2.1. The intersection corresponds to the shaded lens-shaped region that lies within both ovals.

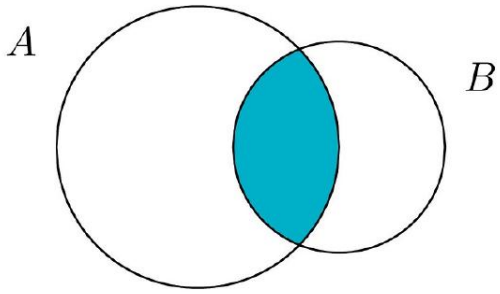


Figure 5.2.1: *The Intersection of Events A and B*

Example 5.2.3

In the experiment of rolling a single die, find the intersection $E \cap T$ of the events E: *"the number rolled is even"* and T: *"the number rolled is greater than two."*

Solution

The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Since the outcomes that are common to $E = \{2, 4, 6\}$ and $T = \{3, 4, 5, 6\}$ are 4 and 6, $E \cap T = \{4, 6\}$.

In words the intersection is described by "the number rolled is even and is greater than two." The only numbers between one and six that are both even and greater than two are four and six, corresponding to $E \cap T$ given above.

Example 5.2.4

A single die is rolled.

1. Suppose the die is fair. Find the probability that the number rolled is both even and greater than two.
2. Suppose the die has been "loaded" so that $P(1) = \frac{1}{12}$, $P(6) = \frac{3}{12}$, and the remaining four outcomes are equally likely with one another. Now find the probability that the number rolled is both even and greater than two.

Solution

In both cases the sample space is $S = \{1, 2, 3, 4, 5, 6\}$ and the event in question is the intersection $E \cap T = \{4, 6\}$ of the previous example.

1. Since the die is fair, all outcomes are equally likely, so by counting we have $P(E \cap T) = \frac{2}{6}$.
2. The information on the probabilities of the six outcomes that we have so far is

$Outcome$	1	2	3	4	5	6
$Probablity$	$\frac{1}{12}$	p	p	p	p	$\frac{3}{12}$

Since $P(1) + P(6) = \frac{4}{6} = \frac{1}{3}$

$$P(2) + P(3) + P(4) + P(5) = 1 - \frac{1}{3} = \frac{2}{3}$$

Thus $4p = \frac{2}{3}$, so $p = \frac{1}{6}$. In particular $P(4) = \frac{1}{6}$ therefore:

$$P(E \cap T) = P(4) + P(6) = \frac{1}{6} + \frac{3}{12} = \frac{5}{12}$$

Definition: mutually exclusive

Events A and B are mutually exclusive (cannot both occur at once) if they have no elements in common.

For A and B to have no outcomes in common means precisely that it is impossible for both A and B to occur on a single trial of the random experiment. This gives the following rule:

Definition: Probability Rule for Mutually Exclusive Events

Events A and B are mutually exclusive if and only if

$$P(A \cap B) = 0$$

Any event A and its complement A^c are mutually exclusive, but A and B can be mutually exclusive without being complements.

Example 5.2.5

In the experiment of rolling a single die, find three choices for an event A so that the events A and E: "the number rolled is even" are mutually exclusive.

Solution

Since $E = \{2, 4, 6\}$ and we want A to have no elements in common with E, any event that does not contain any even number will do. Three choices are $\{1, 3, 5\}$ (the complement E^c, the odds), $\{1, 3\}$, and $\{5\}$.

Union of Events

Definition: Union of Events

The union of events A and B, denoted $A \cup B$, is the collection of all outcomes that are elements of one or the other of the sets A and B, or of both of them. It corresponds to combining descriptions of the two events using the word "or."

To say that the event $A \cup B$ occurred means that on a particular trial of the experiment either A or B occurred (or both did). A visual representation of the union of events A and B in a sample space S is given in Figure 5.2.2. The union corresponds to the shaded region.

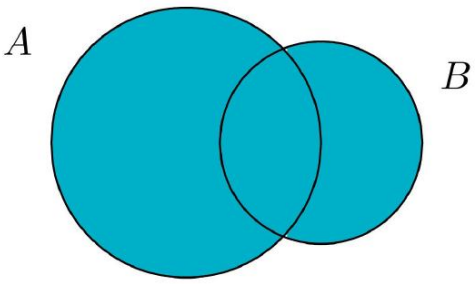


***Figure* 5.2.2**: *The Union of Events A and B*

Example 5.2.6

In the experiment of rolling a single die, find the union of the events E: **"the number rolled is even"** and T: **"the number rolled is greater than two."**

Solution

Since the outcomes that are in either $E=\{2,4,6\}$ or $T=\{3,4,5,6\}$ (or both) are $2,3,4,5$, and 6, that means $E\cup T=\{2,3,4,5,6\}$.

Note that an outcome such as 4 that is in both sets is still listed only once (although strictly speaking it is not incorrect to list it twice).

In words the union is described by "the number rolled is even or is greater than two." Every number between one and six except the number one is either even or is greater than two, corresponding to $E\cup T$ given above.

Example 5.2.7

A two-child family is selected at random. Let B denote the event that at least one child is a boy, let D denote the event that the genders of the two children differ, and let M denote the event that the genders of the two children match. Find $B\cup D$ and $B\cup M$.

Solution

A sample space for this experiment is $S=\{bb,bg,gb,gg\}$, where the first letter denotes the gender of the firstborn child and the second letter denotes the gender of the second child. The events B, D, and M are $B=\{bb,bg,gb\}$, $D=\{bg,gb\}$, $M=\{bb,gg\}$.

Each outcome in D is already in B, so the outcomes that are in at least one or the other of the sets B and D is just the set B itself: $B\cup D=\{bb,bg,gb\}=B$.

Every outcome in the whole sample space S is in at least one or the other of the sets B and M, so $B\cup M=\{bb,bg,gb,gg\}=S$.

Definition: Additive Rule of Probability

A useful property to know is the Additive Rule of Probability, which is

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

The next example, in which we compute the probability of a union both by counting and by using the formula, shows why the last term in the formula is needed.

Example 5.2.8

Two fair dice are thrown. Find the probabilities of the following events:

1. both dice show a four
2. at least one die shows a four

Solution

As was the case with tossing two identical coins, actual experience dictates that for the sample space to have equally likely outcomes we should list outcomes as if we could distinguish the two dice. We could imagine that one of them is red and the other is green. Then any outcome can be labeled as a pair of numbers as in the following display, where the first number in the pair is the number of dots on the top face of the green die and the second number in the pair is the number of dots on the top face of the red die.

$$\begin{array}{cccccc} 1 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{array}$$

1. There are 36 equally likely outcomes, of which exactly one corresponds to two fours, so the probability of a pair of fours is $1/36$.

2. From the table we can see that there are 11 pairs that correspond to the event in question: the six pairs in the fourth row (the green die shows a four) plus the additional five pairs other than the pair 44, already counted, in the fourth column (the red die is four), so the answer is $11/36$. To see how the formula gives the same number, let A_G denote the event that the green die is a four and let A_R denote the event that the red die is a four. Then clearly by counting we get: $P(A_G) = 6/36$ and $P(A_R) = 6/36$. Since $A_G \cap A_R = \{44\}$, $P(A_G \cap A_R) = 1/36$. This is the computation from part 1, of course. Thus by the Additive Rule of Probability we get:

$$P(A_G \cap A_R) = P(A_G) + P(A_R) - P(A_G - A_R) = 6/36 + 6/36 - 1/36 = \frac{11}{36}$$

Example 5.2.9

A tutoring service specializes in preparing adults for high school equivalence tests. Among all the students seeking help from the service, 63% need help in mathematics, 34% need help in English, and 27% need help in both mathematics and English. What is the percentage of students who need help in either mathematics or English?

Solution

Imagine selecting a student at random, that is, in such a way that every student has the same chance of being selected. Let M denote the event "the student needs help in mathematics" and let E denote the event "the student needs help in English." The information given is that $P(M) = 0.63$, $P(E) = 0.34$ and $P(M \cap E) = 0.27$. Thus the Additive Rule of Probability gives:

$$P(M \cup E) = P(M) + P(E) - P(M \cap E) = 0.63 + 0.34 - 0.27 = 0.70$$

Note how the naïve reasoning that if 63% need help in mathematics and 34% need help in English then 63 plus 34 or 97% need help in one or the other gives a number that is too large. The percentage that need help in both subjects must be subtracted off, else the people needing help in both are counted twice, once for needing help in mathematics and once again for needing help in English. The simple sum of the probabilities would work if the events in question were mutually exclusive, for then $P(A \cap B)$ is zero, and makes no difference.

Example 5.2.10

Volunteers for a disaster relief effort were classified according to both specialty (C: construction, E: education, M: medicine) and language ability (S: speaks a single language fluently, T: speaks two or more languages fluently). The results are shown in the following two-way classification table:

Specialty	Language Ability	
	S	T
C	12	1
E	4	3
M	6	2

The first row of numbers means that 12 volunteers whose specialty is construction speak a single language fluently, and 1 volunteer whose specialty is construction speaks at least two languages fluently. Similarly for the other two rows.

A volunteer is selected at random, meaning that each one has an equal chance of being chosen. Find the probability that:

1. his specialty is medicine and he speaks two or more languages;
2. either his specialty is medicine or he speaks two or more languages;
3. his specialty is something other than medicine.

Solution

When information is presented in a two-way classification table it is typically convenient to adjoin to the table the row and column totals, to produce a new table like this:

Specialty	Language Ability		Total
	S	T	
C	12	1	13
E	4	3	7
M	6	2	8
Total	22	6	28

1. The probability sought is $P(M \cap T)$. The table shows that there are 2 such people, out of 28 in all, hence $P(M \cap T) = 2/28 \approx 0.07$ or about a 7% chance.
2. The probability sought is $P(M \cup T)$. The third row total and the grand total in the sample give $P(M) = 8/28$. The second column total and the grand total give $P(T) = 6/28$. Thus using the result from part (1),

$$P(M \cup T) = P(M) + P(T) - P(M \cap T) = 828 + 628 - 228 = 1228 \approx 0.43$$

   or about a 43% chance.

3. This probability can be computed in two ways. Since the event of interest can be viewed as the event $C \cup E$ and the events C and E are mutually exclusive, the answer is, using the first two row totals,

$$P(C \cup E) = P(C) + P(E) - P(C \cap E) = 1328 + 728 - 028 = 2028 \approx 0.71$$

On the other hand, the event of interest can be thought of as the complement M^c of M, hence using the value of $P(M)$computed in part (2),

$$P(M^c) = 1 - P(M) = 1 - 828 = 2028 \approx 0.71$$

as before.

Key Takeaway

- The probability of an event that is a complement or union of events of known probability can be computed using formulas.

5.3: Conditional Probability and Independent Events

Learning Objectives

- To learn the concept of a conditional probability and how to compute it.
- To learn the concept of independence of events, and how to apply it.

Suppose a fair die has been rolled and you are asked to give the probability that it was a five. There are six equally likely outcomes, so your answer is $1/6$. But suppose that before you give your answer you are given the extra information that the number rolled was odd. Since there are only three odd numbers that are possible, one of which is five, you would certai: nly revise your estimate of the likelihood that a five was rolled from $1/6$ to $1/3$. In general, the *revised* probability that an event A has occurred, taking into account the additional information that another event B has definitely occurred on this trial of the experiment, is called the *conditional probability of A given B* and is denoted by $P(A \mid B)$. The reasoning employed in this example can be generalized to yield the computational formula in the following definition.

Definition: conditional probability

The *conditional probability* of A given B, denoted $P(A \mid B)$, is the probability that event A has occurred in a trial of a random experiment for which it is known that event B has definitely occurred. It may be computed by means of the following formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{5.3.1}$$

Example 5.3.1: Rolling a Die

A fair (unbiased) die is rolled.

a. Find the probability that the number rolled is a five, given that it is odd.
b. Find the probability that the number rolled is odd, given that it is a five.

Solution

The sample space for this experiment is the set $S = 1, 2, 3, 4, 5, 6$ consisting of six equally likely outcomes. Let F denote the event "a five is rolled" and let O denote the event "an odd number is rolled," so that

$$F = 5 \text{ and } O = 1, 3, 5$$

a. This is the introductory example, so we already know that the answer is $1/3$. To use Equation 5.3.1 to confirm this we must replace A in the formula (the event whose likelihood we seek to estimate) by F and replace B (the event we know for certain has occurred) by O:

$$P(F \mid O) = \frac{P(F \cap O)}{P(O)}$$

Since

$$F \cap O = 5 \cap 1, 3, 5 = 5, \; P(F \cap O) = 1/6$$

Since

$$O = 1, 3, 5, \; P(O) = 3/6.$$

Thus

$$P(F \mid O) = \frac{P(F \cap O)}{P(O)} = \frac{1/6}{3/6} = \frac{1}{3}$$

b. This is the same problem, but with the roles of F and O reversed. Since we are given that the number that was rolled is five, which is odd, the probability in question must be 1. To apply Equation 5.3.1 to this case we must now replace A (the event whose likelihood we seek to estimate) by O and B (the event we know for certain has occurred) by F:

$$P(O \mid F) = \frac{P(O \cap F)}{P(F)}$$

Obviously $P(F) = 1/6$. In part (a) we found that $P(F \mid O) = 1/6$. Thus

$$P(O \mid F) = \frac{P(O \cap F)}{P(F)} = \frac{1/6}{1/6} = 1$$

Just as we did not need the computational formula in this example, we do not need it when the information is presented in a *two-way classification table*, as in the next example.

Example 5.3.2: Marriage and Gender

In a sample of 902 individuals under 40 who were or had previously been married, each person was classified according to gender and age at first marriage. The results are summarized in the following two-way classification table, where the meaning of the labels is:

- M: male
- F: female
- E: a teenager when first married
- W: in one's twenties when first married
- H: in one's thirties when first married

	E	W	H	Total
M	43	293	114	450
F	82	299	71	452
Total	125	592	185	902

The numbers in the first row mean that 43 people in the sample were men who were first married in their teens, 293 were men who were first married in their twenties, 114 men who were first married in their thirties, and a total of 450 people in the sample were men. Similarly for the numbers in the second row. The numbers in the last row mean that, irrespective of gender, 125 people in the sample were married in their teens, 592 in their twenties, 185 in their thirties, and that there were 902 people in the sample in all. Suppose that the proportions in the sample accurately reflect those in the population of all individuals in the population who are under 40 and who are or have previously been married. Suppose such a person is selected at random.

1. Find the probability that the individual selected was a teenager at first marriage.
2. Find the probability that the individual selected was a teenager at first marriage, given that the person is male.

Solution

It is natural to let E also denote the event that the person selected was a teenager at first marriage and to let M denote the event that the person selected is male.

1. According to the table, the proportion of individuals in the sample who were in their teens at their first marriage is $125/902$ This is the relative frequency of such people in the population, hence $P(E) = 125/902 \approx 0.139$ or about 14%.
2. Since it is known that the person selected is male, all the females may be removed from consideration, so that only the row in the table corresponding to men in the sample applies:

	E	W	H	Total
M	43	293	114	450

The proportion of males in the sample who were in their teens at their first marriage is $43/450$ This is the relative frequency of such people in the population of males, hence $P(E/M) = 43/450 \approx 0.096$ or about 10%.

In the next example, the computational formula in the definition must be used.

Example 5.3.3: Body Weigth and hypertension

Suppose that in an adult population the proportion of people who are both overweight and suffer hypertension is 0.09; the proportion of people who are not overweight but suffer hypertension is 0.11; the proportion of people who are overweight but do not suffer hypertension is 0.02; and the proportion of people who are neither overweight nor suffer hypertension is 0.78. An adult is randomly selected from this population.

a. Find the probability that the person selected suffers hypertension given that he is overweight.
b. Find the probability that the selected person suffers hypertension given that he is not overweight.
c. Compare the two probabilities just found to give an answer to the question as to whether overweight people tend to suffer from hypertension.

Solution:

Let H denote the event "the person selected suffers hypertension." Let O denote the event "the person selected is overweight." The probability information given in the problem may be organized into the following contingency table:

	O	O^c
H	0.09	0.11
H^c	0.02	0.78

a. Using the formula in the definition of conditional probability (Equation 5.3.1),

$$P(H|O) = \frac{P(H \cap O)}{P(O)} = \frac{0.09}{0.09 + 0.02} = 0.8182$$

b. Using the formula in the definition of conditional probability (Equation 5.3.1),

$$P(H|O) = \frac{P(H \cap O^c)}{P(O^c)} = \frac{0.11}{0.11 + 0.78} = 0.1236$$

c. $P(H|O) = 0.8182$ is over six times as large as $P(H|O^c) = 0.1236$, which indicates a much higher rate of hypertension among people who are overweight than among people who are not overweight. It might be interesting to note that a direct comparison of $P(H \cap O) = 0.09$ and $P(H \cap O^c) = 0.11$ does not answer the same question.

Independent Events

Although typically we expect the conditional probability $P(A \mid B)$ to be different from the probability $P(A)$ of A, it does not have to be different from $P(A)$. When $P(A \mid B) = P(A)$, the occurrence of B has no effect on the likelihood of A. Whether or not the event A has occurred is independent of the event B.

Using algebra it can be shown that the equality $P(A \mid B) = P(A)$ holds if and only if the equality $P(A \cap B) = P(A) \cdot P(B)$ holds, which in turn is true if and only if $P(B \mid A) = P(B)$. This is the basis for the following definition.

Definition: Independent and Dependent Events

Events A and B are *independent* (i.e., events whose probability of occurring together is the product of their individual probabilities). if

$$P(A \cap B) = P(A) \cdot P(B)$$

If A and B are not independent then they are *dependent*.

The formula in the definition has two practical but exactly opposite uses:

- In a situation in which we can compute all three probabilities $P(A), P(B)$ and $P(A \cap B)$, it is used to check whether or not the events A and B are independent:
  - If $P(A \cap B) = P(A) \cdot P(B)$, then A and B are independent.
  - If $P(A \cap B) \neq P(A) \cdot P(B)$, then A and B are not independent.
- In a situation in which each of $P(A)$ and $P(B)$ can be computed and it is known that A and B are independent, then we can compute $P(A \cap B)$ by multiplying together $P(A)$ and $P(B)$: $P(A \cap B) = P(A) \cdot P(B)$.

Example 5.3.4: Rolling a Die again

A single fair die is rolled. Let $A = \{3\}$ and $B = \{1, 3, 5\}$. Are A and B independent?

Solution

In this example we can compute all three probabilities $P(A) = 1/6$, $P(B) = 1/2$, and $P(A \cap B) = P(\{3\}) = 1/6$. Since the product $P(A) \cdot P(B) = (1/6)(1/2) = 1/12$ is not the same number as $P(A \cap B) = 1/6$, the events A and B are not independent.

Example 5.3.5

The two-way classification of married or previously married adults under 40 according to gender and age at first marriage produced the table

	E	W	H	Total
M	43	293	114	450
F	82	299	71	452
Total	125	592	185	902

Determine whether or not the events F: “female” and E: “was a teenager at first marriage” are independent.

Solution

The table shows that in the sample of 902 such adults, 452 were female, 125 were teenagers at their first marriage, and 82 were females who were teenagers at their first marriage, so that

$$P(F) = \frac{452}{902},$$

$$P(E) = \frac{125}{902}$$

$$P(F \cap E) = \frac{82}{902}$$

Since

$$\begin{aligned} P(F) \cdot P(E) &= \frac{452}{902} \cdot \frac{125}{902} \\ &= 0.069 \end{aligned}$$

is not the same as

$$P(F \cap E) = \frac{82}{902} = 0.091$$

we conclude that the two events are not independent.

Example 5.3.6

Many diagnostic tests for detecting diseases do not test for the disease directly but for a chemical or biological product of the disease, hence are not perfectly reliable. The *sensitivity* of a test is the probability that the test will be positive when administered to a person who has the disease. The higher the sensitivity, the greater the detection rate and the lower the false negative rate.

Suppose the sensitivity of a diagnostic procedure to test whether a person has a particular disease is 92%. A person who actually has the disease is tested for it using this procedure by two independent laboratories.

a. What is the probability that both test results will be positive?
b. What is the probability that at least one of the two test results will be positive?

Solution

a. Let A_1 denote the event "the test by the first laboratory is positive" and let A_2 denote the event "the test by the second laboratory is positive." Since A_1 and A_2 are independent,

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1) \cdot P(A_2) \\ &= 0.92 \times 0.92 \\ &= 0.8464 \end{aligned}$$

b. Using the Additive Rule for Probability and the probability just computed,

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= 0.92 + 0.92 - 0.8464 \\ &= 0.9936 \end{aligned}$$

Example 5.3.7: specificity of a diagnostic test

The *specificity* of a diagnostic test for a disease is the probability that the test will be negative when administered to a person who does not have the disease. The higher the specificity, the lower the false positive rate. Suppose the specificity of a diagnostic procedure to test whether a person has a particular disease is 89%.

a. A person who does not have the disease is tested for it using this procedure. What is the probability that the test result will be positive?
b. A person who does not have the disease is tested for it by two independent laboratories using this procedure. What is the probability that both test results will be positive?

Solution

a. Let B denote the event "the test result is positive." The complement of B is that the test result is negative, and has probability the specificity of the test, 0.89. Thus

$$P(B) = 1 - P(B^c) = 1 - 0.89 = 0.11$$

b. Let B_1 denote the event "the test by the first laboratory is positive" and let B_2 denote the event "the test by the second laboratory is positive." Since B_1 and B_2 are independent, by part (a) of the example

$$P(B_1 \cap B_2) = P(B_1) \cdot P(B_2) = 0.11 \times 0.11 = 0.0121$$

The concept of independence applies to any number of events. For example, three events A, B, and C are independent if $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$. Note carefully that, as is the case with just two events, this is not a formula that is always valid, but holds precisely when the events in question are independent.

Example 5.3.8: redundancy

The reliability of a system can be enhanced by redundancy, which means building two or more independent devices to do the same job, such as two independent braking systems in an automobile. Suppose a particular species of trained dogs has a 90% chance of detecting contraband in airline luggage. If the luggage is checked three times by three different dogs independently of one another, what is the probability that contraband will be detected?

Solution

Let D_1 denote the event that the contraband is detected by the first dog, D_2 the event that it is detected by the second dog, and D_3 the event that it is detected by the third. Since each dog has a 90% of detecting the contraband, by the Probability Rule for Complements it has a 10% chance of failing. In symbols,

$$P(D_1^c) = 0.10, \quad P(D_2^c) = 0.10, \quad P(D_3^c) = 0.10$$

Let D denote the event that the contraband is detected. We seek $P(D)$. It is easier to find $P(D^c)$, because although there are several ways for the contraband to be detected, there is only one way for it to go undetected: all three dogs must fail. Thus $D^c = D_1^c \cap D_2^c \cap D_3^c$ and

$$P(D) = 1 - P(D^c) = 1 - P(D_1^c \cap D_2^c \cap D_3^c)$$

But the events D_1, D_2, and D_3 are independent, which implies that their complements are independent, so

$$P(D_1^c \cap D_2^c \cap D_3^c) = P(D_1^c) \cdot P(D_2^c) \cdot P(D_3^c) = 0.10 \times 0.10 \times 0.10 = 0.001$$

Using this number in the previous display we obtain

$$P(D) = 1 - 0.001 = 0.999$$

That is, although any one dog has only a 90% chance of detecting the contraband, three dogs working independently have a 99.9% chance of detecting it.

Probabilities on Tree Diagrams

Some probability problems are made much simpler when approached using a tree diagram. The next example illustrates how to place probabilities on a tree diagram and use it to solve a problem.

Example 5.3.9: A jar of Marbles

A jar contains 10 marbles, 7 black and 3 white. Two marbles are drawn without replacement, which means that the first one is not put back before the second one is drawn.

a. What is the probability that both marbles are black?
b. What is the probability that exactly one marble is black?
c. What is the probability that at least one marble is black?

Solution

A tree diagram for the situation of drawing one marble after the other without replacement is shown in Figure 5.3.1. The circle and rectangle will be explained later, and should be ignored for now.

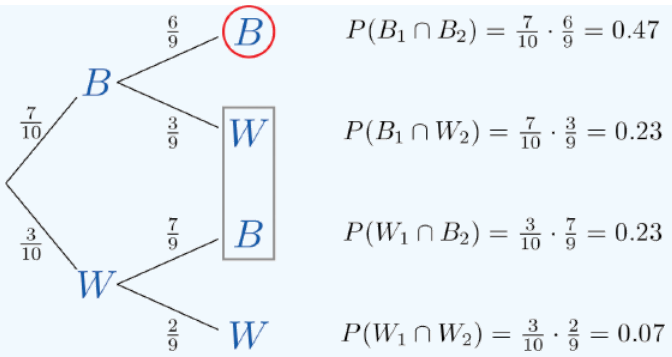


Figure 5.3.1: Tree Diagram for Drawing Two Marbles

The numbers on the two leftmost branches are the probabilities of getting either a black marble, 7 out of 10, or a white marble, 3 out of 10, on the first draw. The number on each remaining branch is the probability of the event corresponding to the node on the right end of the branch occurring, given that the event corresponding to the node on the left end of the branch has occurred. Thus for the top branch, connecting the two Bs, it is $P(B_2 \mid B_1)$, where B_1 denotes the event "the first marble drawn is black" and B_2 denotes the event "the second marble drawn is black." Since after drawing a black marble out there are 9 marbles left, of which 6 are black, this probability is $6/9$.

The number to the right of each final node is computed as shown, using the principle that if the formula in the Conditional Rule for Probability is multiplied by $P(B)$, then the result is

$$P(B \cap A) = P(B) \cdot P(A \mid B)$$

a. The event "both marbles are black" is $B_1 \cap B_2$ and corresponds to the top right node in the tree, which has been circled. Thus as indicated there, it is 0.47.
b. The event "exactly one marble is black" corresponds to the two nodes of the tree enclosed by the rectangle. The events that correspond to these two nodes are mutually exclusive: black followed by white is incompatible with white followed by black. Thus in accordance with the Additive Rule for Probability we merely add the two probabilities next to these nodes, since what would be subtracted from the sum is zero. Thus the probability of drawing exactly one black marble in two tries is $0.23 + 0.23 = 0.46$.
c. The event "at least one marble is black" corresponds to the three nodes of the tree enclosed by either the circle or the rectangle. The events that correspond to these nodes are mutually exclusive, so as in part (b) we merely add the probabilities next to these nodes. Thus the probability of drawing at least one black marble in two tries is $0.47 + 0.23 + 0.23 = 0.93$.

Of course, this answer could have been found more easily using the Probability Law for Complements, simply subtracting the probability of the complementary event, "two white marbles are drawn," from 1 to obtain $1 - 0.07 = 0.93$.

As this example shows, finding the probability for each branch is fairly straightforward, since we compute it knowing everything that has happened in the sequence of steps so far. Two principles that are true in general emerge from this example:

Probabilities on Tree Diagrams

- The probability of the event corresponding to any node on a tree is the product of the numbers on the unique path of branches that leads to that node from the start.
- If an event corresponds to several final nodes, then its probability is obtained by *adding* the numbers next to those nodes.

Key Takeaway

- A conditional probability is the probability that an event has occurred, taking into account additional information about the result of the experiment.
- A conditional probability can always be computed using the formula in the definition. Sometimes it can be computed by discarding part of the sample space.
- Two events A and B are independent if the probability $P(A \cap B)$ of their intersection $A \cap B$ is equal to the product $P(A) \cdot P(B)$ of their individual probabilities.

5.E: Basic Concepts of Probability (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

3.1: Sample Spaces, Events, and Their Probabilities

Basic

Q3.1.1

A box contains 10 white and 10 black marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, two marbles in succession and noting the color each time. (To draw “with replacement” means that the first marble is put back before the second marble is drawn.)

Q3.1.2

A box contains 16 white and 16 black marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, three marbles in succession and noting the color each time. (To draw “with replacement” means that each marble is put back before the next marble is drawn.)

Q3.1.3

A box contains 8 red, 8 yellow, and 8 green marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, two marbles in succession and noting the color each time.

Q3.1.4

A box contains 6 red, 6 yellow, and 6 green marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, three marbles in succession and noting the color each time.

Q3.1.5

In the situation of **Exercise 1**, list the outcomes that comprise each of the following events.

a. At least one marble of each color is drawn.
b. No white marble is drawn.

Q3.1.6

In the situation of **Exercise 2**, list the outcomes that comprise each of the following events.

a. At least one marble of each color is drawn.
b. No white marble is drawn.
c. More black than white marbles are drawn.

Q3.1.7

In the situation of **Exercise 3**, list the outcomes that comprise each of the following events.

a. No yellow marble is drawn.
b. The two marbles drawn have the same color.
c. At least one marble of each color is drawn.

Q3.1.8

In the situation of **Exercise 4**, list the outcomes that comprise each of the following events.

a. No yellow marble is drawn.
b. The three marbles drawn have the same color.
c. At least one marble of each color is drawn.

Q3.1.9

Assuming that each outcome is equally likely, find the probability of each event in **Exercise 5**.

Q3.1.10

Assuming that each outcome is equally likely, find the probability of each event in **Exercise 6**.

Q3.1.11

Assuming that each outcome is equally likely, find the probability of each event in **Exercise 7**.

Q3.1.12

Assuming that each outcome is equally likely, find the probability of each event in **Exercise 8**.

Q3.1.13

A sample space is $S = \{a, b, c, d, e\}$. Identify two events as $U = \{a, b, d\}$ and $V = \{b, c, d\}$. Suppose $P(a)$ and $P(b)$ are each 0.2 and $P(c)$ and $P(d)$ are each 0.1.

a. Determine what $P(e)$ must be.
b. Find $P(U)$.
c. Find $P(V)$

Q3.1.14

A sample space is $S = \{u, v, w, x\}$. Identify two events as $A = \{v, w\}$ and $B = \{u, w, x\}$. Suppose $P(u) = 0.22$, $P(w) = 0.36$, and $P(x) = 0.27$.

a. Determine what $P(v)$ must be.
b. Find $P(A)$.
c. Find $P(B)$.

Q3.1.15

A sample space is $S = \{m, n, q, r, s\}$. Identify two events as $U = \{m, q, s\}$ and $V = \{n, q, r\}$. The probabilities of some of the outcomes are given by the following table:

$Outcome$	m	n	q	r	s
$Probability$	0.18	0.16		0.24	0.21

(5.E.1)

a. Determine what $P(q)$ must be.
b. Find $P(U)$.
c. Find $P(V)$.

Q3.1.16

A sample space is $S = \{d, e, f, g, h\}$. Identify two events as $M = \{e, f, g, h\}$ and $N = \{d, g\}$. The probabilities of some of the outcomes are given by the following table:

$Outcome$	d	e	f	g	h
$Probability$	0.22	0.13	0.27		0.19

(5.E.2)

a. Determine what $P(g)$ must be.
b. Find $P(M)$.
c. Find $P(N)$.

Applications

Q3.1.17

The sample space that describes all three-child families according to the genders of the children with respect to birth order was constructed in "Example 3.1.4". Identify the outcomes that comprise each of the following events in the experiment of selecting a three-child family at random.

a. At least one child is a girl.
b. At most one child is a girl.
c. All of the children are girls.

d. Exactly two of the children are girls.
e. The first born is a girl.

Q3.1.18

The sample space that describes three tosses of a coin is the same as the one constructed in "Example 3.1.4" with “boy” replaced by “heads” and “girl” replaced by “tails.” Identify the outcomes that comprise each of the following events in the experiment of tossing a coin three times.

a. The coin lands heads more often than tails.
b. The coin lands heads the same number of times as it lands tails.
c. The coin lands heads at least twice.
d. The coin lands heads on the last toss.

Q3.1.19

Assuming that the outcomes are equally likely, find the probability of each event in **Exercise 17**.

Q3.1.20

Assuming that the outcomes are equally likely, find the probability of each event in **Exercise 18**.

Additional Exercises

Q3.1.21

The following two-way contingency table gives the breakdown of the population in a particular locale according to age and tobacco usage:

Age	Tobacco Use	
	Smoker	Non-smoker
Under 30	0.05	0.20
Over 30	0.20	0.55

A person is selected at random. Find the probability of each of the following events.

a. The person is a smoker.
b. The person is under 30.
c. The person is a smoker who is under 30.

Q3.1.22

The following two-way contingency table gives the breakdown of the population in a particular locale according to party affiliation ($A, B, C,$ or None) and opinion on a bond issue:

Affiliation	Opinion		
	Favors	Opposes	Undecided
A	0.12	0.09	0.07
B	0.16	0.12	0.14
C	0.04	0.03	0.06
None	0.08	0.06	0.03

A person is selected at random. Find the probability of each of the following events.

a. The person is affiliated with party B.
b. The person is affiliated with some party.
c. The person is in favor of the bond issue.

d. The person has no party affiliation and is undecided about the bond issue.

Q3.1.23

The following two-way contingency table gives the breakdown of the population of married or previously married women beyond child-bearing age in a particular locale according to age at first marriage and number of children:

Age	Number of Children		
	0	1 *or* 2	3 or More
Under 20	0.02	0.14	0.08
20 – 29	0.07	0.37	0.11
30 and above	0.10	0.10	0.01

A woman is selected at random. Find the probability of each of the following events.

a. The woman was in her twenties at her first marriage.
b. The woman was 20 or older at her first marriage.
c. The woman had no children.
d. The woman was in her twenties at her first marriage and had at least three children.

Q3.1.24

The following two-way contingency table gives the breakdown of the population of adults in a particular locale according to highest level of education and whether or not the individual regularly takes dietary supplements:

Education	Use of Supplements	
	Takes	Does Not Take
No High School Diploma	0.04	0.06
High School Diploma	0.06	0.44
Undergraduate Degree	0.09	0.28
Graduate Degree	0.01	0.02

An adult is selected at random. Find the probability of each of the following events.

a. The person has a high school diploma and takes dietary supplements regularly.
b. The person has an undergraduate degree and takes dietary supplements regularly.
c. The person takes dietary supplements regularly.
d. The person does not take dietary supplements regularly.

Large Data Set Exercises

Q3.1.25

Large Data Set 4 and Data Set 4A record the results of 500 tosses of a coin. Find the relative frequency of each outcome $1, 2, 3, 4, 5,$ *and* 6 Does the coin appear to be "balanced" or "fair"?

Q3.1.26

Large Data Set 6, Data Set 6A, and Data Set 6B record results of a random survey of 200 voters in each of two regions, in which they were asked to express whether they prefer Candidate A for a U.S. Senate seat or prefer some other candidate.

a. Find the probability that a randomly selected voter among these 400 prefers Candidate A.
b. Find the probability that a randomly selected voter among the 200 who live in Region 1 prefers Candidate A (separately recorded in Large Data Set 6A).

c. Find the probability that a randomly selected voter among the 200 who live in Region 2 prefers Candidate A (separately recorded in Large Data Set 6B).

Answers

S3.1.1

$S = \{bb, bw, wb, ww\}$

S3.1.3

$S = \{rr, ry, rg, yr, yy, yg, gr, gy, gg\}$

S3.1.5

a. $\{bw, wb\}$
b. $\{bb\}$

S3.1.7

a. $\{rr, rg, gr, gg\}$
b. $\{rr, yy, gg\}$
c. $\varnothing$

S3.1.9

a. $1/4$
b. $2/4$

S3.1.11

a. $4/9$
b. $3/9$
c. 0

S3.1.13

a. 0.4
b. 0.5
c. 0.4

S3.1.15

a. 0.61
b. 0.6
c. 0.21

S3.1.17

a. $\{gbb, gbg, ggb, ggg\}$
b. $\{bgg, gbg, ggb\}$
c. $\{ggg\}$
d. $\{bbb, bbg, bgb, gbb\}$
e. $\{bbg, bgb, bgg, gbb, gbg, ggb, ggg\}$

S3.1.19

a. $4/8$
b. $3/8$
c. $1/8$
d. $4/8$
e. $7/8$

S3.1.21

a. 0.05
b. 0.25
c. 0.25

S3.1.23

a. 0.11
b. 0.19
c. 0.76
d. 0.55

S3.1.25

The relative frequencies for 1 through 6 are $0.16, 0.194, 0.162, 0.164, 0.154$ *and* 0.166 It would appear that the die is not balanced.

3.2: Complements, Intersections and Unions

Basic

1. For the sample space $S = \{a, b, c, d, e\}$ identify the complement of each event given.
   a. $A = \{a, d, e\}$
   b. $B = \{b, c, d, e\}$
   c. S
2. For the sample space $S = \{r, s, t, u, v\}$ identify the complement of each event given.
   a. $R = \{t, u\}$
   b. $T = \{r\}$
   c. $\varnothing$ (the "empty" set that has no elements)
3. The sample space for three tosses of a coin is $S = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$ Define events

$$\begin{array}{c} \text{H:at least one head is observed} \\ \text{M:more heads than tails are observed} \end{array} \tag{5.E.3}$$

   a. List the outcomes that comprise H and M.
   b. List the outcomes that comprise $H \cap M$, $H \cup M$, and H^c.
   c. Assuming all outcomes are equally likely, find $P(H \cap M)$, $P(H \cup M)$, and $P(H^c)$.
   d. Determine whether or not H^c and M are mutually exclusive. Explain why or why not.
4. For the experiment of rolling a single six-sided die once, define events

$$\begin{array}{c} \text{T:the number rolled is three} \\ \text{G:the number rolled is four or greater} \end{array} \tag{5.E.4}$$

   a. List the outcomes that comprise T and G.
   b. List the outcomes that comprise $T \cap G$, $T \cup G$, T^c, and $(T \cup G)^c$.
   c. Assuming all outcomes are equally likely, find $P(T \cap G)$, $P(T \cup G)$, and $P(T^c)$.
   d. Determine whether or not T and G are mutually exclusive. Explain why or why not.
5. A special deck of 16 cards has 4 that are blue, 4 yellow, 4 green, and 4 red. The four cards of each color are numbered from one to four. A single card is drawn at random. Define events

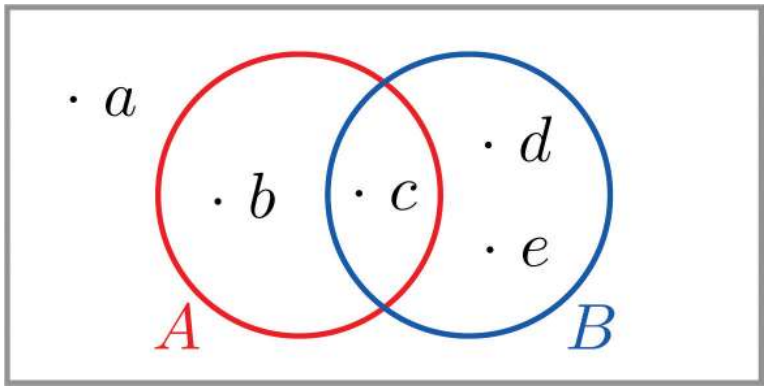
$$\begin{array}{c} \text{B:the card is blue} \\ \text{R:the card is red} \\ \text{N:the number on the card is at most two} \end{array} \tag{5.E.5}$$

   a. List the outcomes that comprise B, R, and N.
   b. List the outcomes that comprise $B \cap R$, $B \cup R$, $B \cap N$, $R \cup N$, B^c, and $(B \cup R)^c$.
   c. Assuming all outcomes are equally likely, find the probabilities of the events in the previous part.
   d. Determine whether or not B and N are mutually exclusive. Explain why or why not.
6. In the context of the previous problem, define events

Y:the card is yellow (5.E.6)
I:the number on the card is not a one
J:the number on the card is a two or a four

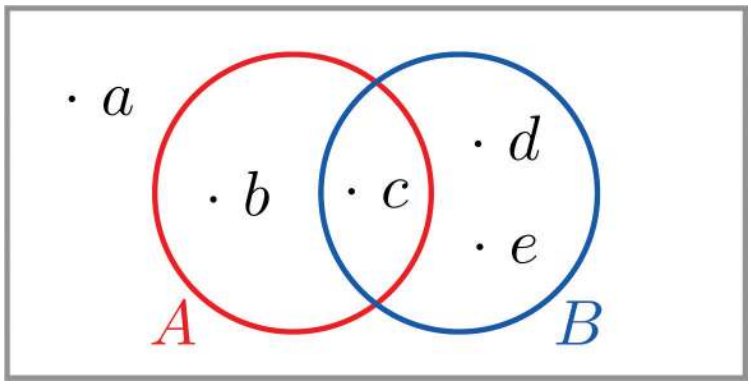
a. List the outcomes that comprise Y, I, and J.
b. List the outcomes that comprise $Y \cap I$, $Y \cup J$, $I \cap J$, I^c, and $(Y \cup J)^c$.
c. Assuming all outcomes are equally likely, find the probabilities of the events in the previous part.
d. Determine whether or not I^c and J are mutually exclusive. Explain why or why not.

7. The Venn diagram provided shows a sample space and two events A and B. Suppose $P(a) = 0.13, P(b) = 0.09, P(c) = 0.27, P(d) = 0.20,$ and $P(e) = 0.31$. Confirm that the probabilities of the outcomes add up to 1, then compute the following probabilities.



a. $P(A)$.
b. $P(B)$.
c. $P(A^c)$. Two ways: (i) by finding the outcomes in A^c and adding their probabilities, and (ii) using the Probability Rule for Complements.
d. $P(A \cap B)$.
e. $P(A \cup B)$ Two ways: (i) by finding the outcomes in $A \cup B$ and adding their probabilities, and (ii) using the Additive Rule of Probability.

8. The Venn diagram provided shows a sample space and two events A and B. Suppose $P(a) = 0.32, P(b) = 0.17, P(c) = 0.28,$ and $P(d) = 0.23$. Confirm that the probabilities of the outcomes add up to 1, then compute the following probabilities.



a. $P(A)$.
b. $P(B)$.
c. $P(A^c)$. Two ways: (i) by finding the outcomes in A^c and adding their probabilities, and (ii) using the Probability Rule for Complements.
d. $P(A \cap B)$.
e. $P(A \cup B)$ Two ways: (i) by finding the outcomes in $A \cup B$ and adding their probabilities, and (ii) using the Additive Rule of Probability.

9. Confirm that the probabilities in the two-way contingency table add up to 1, then use it to find the probabilities of the events indicated.

	U	V	W
A	0.15	0.00	0.23
B	0.22	0.30	0.10

a. $P(A), P(B), P(A \cap B)$.
b. $P(U), P(W), P(U \cap W)$.
c. $P(U \cup W)$.
d. $P(V^c)$.
e. Determine whether or not the events A and U are mutually exclusive; the events A and V.

10. Confirm that the probabilities in the two-way contingency table add up to 1, then use it to find the probabilities of the events indicated.

	R	S	T
M	0.09	0.25	0.19
N	0.31	0.16	0.00

a. $P(R), P(S), P(R \cap S)$.
b. $P(M), P(N), P(M \cap N)$.
c. $P(R \cup S)$.
d. $P(R^c)$.
e. Determine whether or not the events N and S are mutually exclusive; the events N and T.

Applications

11. Make a statement in ordinary English that describes the complement of each event (do not simply insert the word "not").
a. In the roll of a die: "five or more."
b. In a roll of a die: "an even number."
c. In two tosses of a coin: "at least one heads."
d. In the random selection of a college student: "Not a freshman."

12. Make a statement in ordinary English that describes the complement of each event (do not simply insert the word "not").
a. In the roll of a die: "two or less."
b. In the roll of a die: "one, three, or four."
c. In two tosses of a coin: "at most one heads."
d. In the random selection of a college student: "Neither a freshman nor a senior."

13. The sample space that describes all three-child families according to the genders of the children with respect to birth order is $S = \{bbb, bbg, bgb, bgg, gbb, gbg, ggb, ggg\}$. For each of the following events in the experiment of selecting a three-child family at random, state the complement of the event in the simplest possible terms, then find the outcomes that comprise the event and its complement.
a. At least one child is a girl.
b. At most one child is a girl.
c. All of the children are girls.
d. Exactly two of the children are girls.
e. The first born is a girl.

14. The sample space that describes the two-way classification of citizens according to gender and opinion on a political issue is $S = \{mf, ma, mn, ff, fa, fn\}$, where the first letter denotes gender (m: male, f: female) and the second opinion (f: for, a: against, n: neutral). For each of the following events in the experiment of selecting a citizen at random, state the complement of the event in the simplest possible terms, then find the outcomes that comprise the event and its complement.

a. The person is male.
b. The person is not in favor.
c. The person is either male or in favor.
d. The person is female and neutral.

15. A tourist who speaks English and German but no other language visits a region of Slovenia. If 35% of the residents speak English, 15% speak German, and 3% speak both English and German, what is the probability that the tourist will be able to talk with a randomly encountered resident of the region?
16. In a certain country 43% of all automobiles have airbags, 27% have anti-lock brakes, and 13% have both. What is the probability that a randomly selected vehicle will have both airbags and anti-lock brakes?
17. A manufacturer examines its records over the last year on a component part received from outside suppliers. The breakdown on source (supplier A, supplier B) and quality (H: high, U: usable, D: defective) is shown in the two-way contingency table.

	H	U	D
A	0.6937	0.0049	0.0014
B	0.2982	0.0009	0.0009

The record of a part is selected at random. Find the probability of each of the following events.

a. The part was defective.
b. The part was either of high quality or was at least usable, in two ways: (i) by adding numbers in the table, and (ii) using the answer to (a) and the Probability Rule for Complements.
c. The part was defective and came from supplier B.
d. The part was defective or came from supplier B, in two ways: by finding the cells in the table that correspond to this event and adding their probabilities, and (ii) using the Additive Rule of Probability.

18. Individuals with a particular medical condition were classified according to the presence (T) or absence (N) of a potential toxin in their blood and the onset of the condition (E: early, M: midrange, L: late). The breakdown according to this classification is shown in the two-way contingency table.

	E	M	L
T	0.012	0.124	0.013
N	0.170	0.638	0.043

One of these individuals is selected at random. Find the probability of each of the following events.

a. The person experienced early onset of the condition.
b. The onset of the condition was either midrange or late, in two ways: (i) by adding numbers in the table, and (ii) using the answer to (a) and the Probability Rule for Complements.
c. The toxin is present in the person's blood.
d. The person experienced early onset of the condition and the toxin is present in the person's blood.
e. The person experienced early onset of the condition or the toxin is present in the person's blood, in two ways: (i) by finding the cells in the table that correspond to this event and adding their probabilities, and (ii) using the Additive Rule of Probability.

19. The breakdown of the students enrolled in a university course by class (F: freshman, So: sophomore, J: junior, Se: senior) and academic major (S: science, mathematics, or engineering, L: liberal arts, O: other) is shown in the two-way classification table.

Major	Class			
	F	So	J	Se
S	92	42	20	13
L	368	167	80	53

Major	Class			
	F	So	J	Se
O	460	209	100	67

A student enrolled in the course is selected at random. Adjoin the row and column totals to the table and use the expanded table to find the probability of each of the following events.

a. The student is a freshman.
b. The student is a liberal arts major.
c. The student is a freshman liberal arts major.
d. The student is either a freshman or a liberal arts major.
e. The student is not a liberal arts major.

20. The table relates the response to a fund-raising appeal by a college to its alumni to the number of years since graduation.

Response	Years Since Graduation			
	$0-5$	$6-20$	$21-35$	Over 35
Positive	120	440	210	90
None	1380	3560	3290	910

An alumnus is selected at random. Adjoin the row and column totals to the table and use the expanded table to find the probability of each of the following events.

a. The alumnus responded.
b. The alumnus did not respond.
c. The alumnus graduated at least 21 years ago.
d. The alumnus graduated at least 21 years ago and responded.

Additional Exercises

21. The sample space for tossing three coins is $S=\{hhh, hht, hth, htt, thh, tht, tth, ttt\}$

a. List the outcomes that correspond to the statement “All the coins are heads.”
b. List the outcomes that correspond to the statement “Not all the coins are heads.”
c. List the outcomes that correspond to the statement “All the coins are not heads.”

Answers

1. a. $\{b, c\}$
   b. $\{a\}$
   c. $\varnothing$
2.
3. a. $H=\{hhh, hht, hth, htt, thh, tht, tth\},\ M=\{hhh, hht, hth, thh\}$
   b. $H \cap M=\{hhh, hht, hth, thh\}, H \cup M=H, H^{c}=\{ttt\}$
   c. $P(H \cap M)=4/8, P(H \cup M)=7/8, P(H^{c})=1/8$
   d. Mutually exclusive because they have no elements in common.
4.
5. a. $B=\{b1, b2, b3, b4\},\ R=\{r1, r2, r3, r4\},\ N=\{b1, b2, y1, y2, g1, g2, r1, r2\}$
   b. $B \cap R=\varnothing, B \cup R=\{b1, b2, b3, b4, r1, r2, r3, r4\},\ B \cap N=\{b1, b2\},$
   $R \cup N=\{b1, b2, y1, y2, g1, g2, r1, r2, r3, r4\},$
   $B^{c}=\{y1, y2, y3, y4, g1, g2, g3, g4, r1, r2, r3, r4\},\ (B \cup R)^{c}=\{y1, y2, y3, y4, g1, g2, g3, g4\}$
   c. $P(B \cap R)=0,\ P(B \cup R)=8/16,\ P(B \cap N)=2/16,\ P(R \cup N)=10/16,\ P(B^{c})=12/16,\ P((B \cup R)^{c})=8/16$
   d. Not mutually exclusive because they have an element in common.

6.

7. a. 0.36
   b. 0.78
   c. 0.64
   d. 0.27
   e. 0.87

8.

9. a. $P(A) = 0.38,\ P(B) = 0.62,\ P(A \cap B) = 0$
   b. $P(U) = 0.37,\ P(W) = 0.33,\ P(U \cap W) = 0$
   c. 0.7
   d. 0.7
   e. A and U are not mutually exclusive because $P(A \cap U)$ is the nonzero number 0.15. A and V are mutually exclusive because $P(A \cap V) = 0$.

10.

11. a. “four or less”
    b. “an odd number”
    c. “no heads” or “all tails”
    d. “a freshman”

12.

13. a. “All the children are boys.” Event: $\{bbg, bgb, bgg, gbb, gbg, ggb, ggg\}$, Complement: $\{bbb\}$
    b. “At least two of the children are girls” or “There are two or three girls.” Event: $\{bbb, bbg, bgb, gbb\}$, Complement: $\{bgg, gbg, ggb, ggg\}$
    c. “At least one child is a boy.” Event: $\{ggg\}$, Complement: $\{bbb, bbg, bgb, bgg, gbb, gbg, ggb\}$
    d. “There are either no girls, exactly one girl, or three girls.” Event: $\{bgg, gbg, ggb\}$, Complement: $\{bbb, bbg, bgb, gbb, ggg\}$
    e. “The first born is a boy.” Event: $\{gbb, gbg, ggb, ggg\}$, Complement: $\{bbb, bbg, bgb, bgg\}$

14.

15. 0.47

16.

17. a. 0.0023
    b. 0.9977
    c. 0.0009
    d. 0.3014

18.

19. a. $920/1671$
    b. $668/1671$
    c. $368/1671$
    d. $1220/1671$
    e. $1003/1671$

20.

21. a. $\{hhh\}$
    b. $\{hht, hth, htt, thh, tht, tth, ttt\}$
    c. $\{ttt\}$

3.3: Conditional Probability and Independent Events

Basic

1. Q3.3.1For two events A and B, $P(A) = 0.73,\ P(B) = 0.48$ and $P(A \cap B) = 0.29$.
   a. Find $P(A \mid B)$.
   b. Find $P(B \mid A)$.
   c. Determine whether or not A and B are independent.

2. Q3.3.1For two events A and B, $P(A) = 0.26,\ P(B) = 0.37$ and $P(A \cap B) = 0.11$.
   a. Find $P(A \mid B)$.
   b. Find $P(B \mid A)$.
   c. Determine whether or not A and B are independent.
3. Q3.3.1For independent events A and B, $P(A) = 0.81$ and $P(B) = 0.27$.
   a. $P(A \cap B)$.
   b. Find $P(A \mid B)$.
   c. Find $P(B \mid A)$.
4. Q3.3.1For independent events A and B, $P(A) = 0.68$ and $P(B) = 0.37$.
   a. $P(A \cap B)$.
   b. Find $P(A \mid B)$.
   c. Find $P(B \mid A)$.
5. Q3.3.1For mutually exclusive events A and B, $P(A) = 0.17$ and $P(B) = 0.32$.
   a. Find $P(A \mid B)$.
   b. Find $P(B \mid A)$.
6. Q3.3.1For mutually exclusive events A and B, $P(A) = 0.45$ and $P(B) = 0.09$.
   a. Find $P(A \mid B)$.
   b. Find $P(B \mid A)$.
7. Q3.3.1Compute the following probabilities in connection with the roll of a single fair die.
   a. The probability that the roll is even.
   b. The probability that the roll is even, given that it is not a two.
   c. The probability that the roll is even, given that it is not a one.
8. Q3.3.1Compute the following probabilities in connection with two tosses of a fair coin.
   a. The probability that the second toss is heads.
   b. The probability that the second toss is heads, given that the first toss is heads.
   c. The probability that the second toss is heads, given that at least one of the two tosses is heads.
9. Q3.3.1A special deck of 16 cards has 4 that are blue, 4 yellow, 4 green, and 4 red. The four cards of each color are numbered from one to four. A single card is drawn at random. Find the following probabilities.
   a. The probability that the card drawn is red.
   b. The probability that the card is red, given that it is not green.
   c. The probability that the card is red, given that it is neither red nor yellow.
   d. The probability that the card is red, given that it is not a four.
10. Q3.3.1A special deck of 16 cards has 4 that are blue, 4 yellow, 4 green, and 4 red. The four cards of each color are numbered from one to four. A single card is drawn at random. Find the following probabilities.
    a. The probability that the card drawn is a two or a four.
    b. The probability that the card is a two or a four, given that it is not a one.
    c. The probability that the card is a two or a four, given that it is either a two or a three.
    d. The probability that the card is a two or a four, given that it is red or green.
11. Q3.3.1A random experiment gave rise to the two-way contingency table shown. Use it to compute the probabilities indicated.

	R	S
A	0.12	0.18
B	0.28	0.42

    a. $P(A),\ P(R),\ P(A \cap B)$.
    b. Based on the answer to (a), determine whether or not the events A and R are independent.
    c. Based on the answer to (b), determine whether or not $P(A \mid R)$ can be predicted without any computation. If so, make the prediction. In any case, compute $P(A \mid R)$ using the Rule for Conditional Probability.

12. Q3.3.1A random experiment gave rise to the two-way contingency table shown. Use it to compute the probabilities indicated.

	R	S
A	0.13	0.07
B	0.61	0.19

    a. $P(A),\ P(R),\ P(A \cap B)$.
    b. Based on the answer to (a), determine whether or not the events A and R are independent.
    c. Based on the answer to (b), determine whether or not $P(A \mid R)$ can be predicted without any computation. If so, make the prediction. In any case, compute $P(A \mid R)$ using the Rule for Conditional Probability.

13. Q3.3.1Suppose for events A and B in a random experiment $P(A) = 0.70$ and $P(B) = 0.30$.Compute the indicated probability, or explain why there is not enough information to do so.
    a. $P(A \cap B)$.
    b. $P(A \cap B)$, with the extra information that A and B are independent.
    c. $P(A \cap B)$, with the extra information that A and B are mutually exclusive.

14. Q3.3.1Suppose for events A and B in a random experiment $P(A) = 0.50$ and $P(B) = 0.50$. Compute the indicated probability, or explain why there is not enough information to do so.
    a. $P(A \cap B)$.
    b. $P(A \cap B)$, with the extra information that A and B are independent.
    c. $P(A \cap B)$, with the extra information that A and B are mutually exclusive.

15. Q3.3.1Suppose for events $A,\ B,\ and\ C$ connected to some random experiment, $A,\ B,\ and\ C$ are independent and $P(A) = 0.50$, $P(B) = 0.50$ and $P(C) = 0.44$. Compute the indicated probability, or explain why there is not enough information to do so.
    a. $P(A \cap B \cap C)$.
    b. $P(A^c \cap B^c \cap C^c)$.

16. Q3.3.1Suppose for events $A,\ B,\ and\ C$ connected to some random experiment, $A,\ B,\ and\ C$ are independent and $P(A) = 0.95$, $P(B) = 0.73$ and $P(C) = 0.62$. Compute the indicated probability, or explain why there is not enough information to do so.
    a. $P(A \cap B \cap C)$.
    b. $P(A^c \cap B^c \cap C^c)$.

Applications

Q3.3.17

The sample space that describes all three-child families according to the genders of the children with respect to birth order is

$$S = \{bbb, bbg, bgb, bgg, gbb, gbg, ggb, ggg\} \tag{5.E.7}$$

In the experiment of selecting a three-child family at random, compute each of the following probabilities, assuming all outcomes are equally likely.

a. The probability that the family has at least two boys.
b. The probability that the family has at least two boys, given that not all of the children are girls.
c. The probability that at least one child is a boy.
d. The probability that at least one child is a boy, given that the first born is a girl.

Q3.3.18

The following two-way contingency table gives the breakdown of the population in a particular locale according to age and number of vehicular moving violations in the past three years:

Age	Violations		
	0	1	2+

Age	Violations		
	0	1	2+
Under 21	0.04	0.06	0.02
21 – 40	0.25	0.16	0.01
41 – 60	0.23	0.10	0.02
60+	0.08	0.03	0.00

A person is selected at random. Find the following probabilities.

a. The person is under 21.
b. The person has had at least two violations in the past three years.
c. The person has had at least two violations in the past three years, given that he is under 21.
d. The person is under 21, given that he has had at least two violations in the past three years.
e. Determine whether the events "the person is under 21" and "the person has had at least two violations in the past three years" are independent or not.

Q3.3.19

The following two-way contingency table gives the breakdown of the population in a particular locale according to party affiliation (A, B, C, or None) and opinion on a bond issue:

Affiliation	Opinion		
	Favors	Opposes	Undecided
A	0.12	0.09	0.07
B	0.16	0.12	0.14
C	0.04	0.03	0.06
None	0.08	0.06	0.03

A person is selected at random. Find each of the following probabilities.

a. The person is in favor of the bond issue.
b. The person is in favor of the bond issue, given that he is affiliated with party A.
c. The person is in favor of the bond issue, given that he is affiliated with party B.

Q3.3.20

The following two-way contingency table gives the breakdown of the population of patrons at a grocery store according to the number of items purchased and whether or not the patron made an impulse purchase at the checkout counter:

Number of Items	Impulse Purchase	
	Made	Not Made
Few	0.01	0.19
Many	0.04	0.76

A patron is selected at random. Find each of the following probabilities.

a. The patron made an impulse purchase.
b. The patron made an impulse purchase, given that the total number of items purchased was many.
c. Determine whether or not the events "few purchases" and "made an impulse purchase at the checkout counter" are independent.

Q3.3.21

The following two-way contingency table gives the breakdown of the population of adults in a particular locale according to employment type and level of life insurance:

Employment Type	Level of Insurance		
	Low	Medium	High
Unskilled	0.07	0.19	0.00
Semi-skilled	0.04	0.28	0.08
Skilled	0.03	0.18	0.05
Professional	0.01	0.05	0.02

An adult is selected at random. Find each of the following probabilities.

a. The person has a high level of life insurance.
b. The person has a high level of life insurance, given that he does not have a professional position.
c. The person has a high level of life insurance, given that he has a professional position.
d. Determine whether or not the events “has a high level of life insurance” and “has a professional position” are independent.

Q3.3.22

The sample space of equally likely outcomes for the experiment of rolling two fair dice is

$$\begin{array}{cccccc} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{array} \tag{5.E.8}$$

Identify the events N: the sum is at least nine, T: at least one of the dice is a two, and F: at least one of the dice is a five .

a. Find $P(N)$.
b. Find $P(N \mid F)$.
c. Find $P(N \mid T)$.
d. Determine from the previous answers whether or not the events N and F are independent; whether or not N and T are.

Q3.3.23

The *sensitivity* of a drug test is the probability that the test will be positive when administered to a person who has actually taken the drug. Suppose that there are two independent tests to detect the presence of a certain type of banned drugs in athletes. One has sensitivity 0.75; the other has sensitivity 0.85. If both are applied to an athlete who has taken this type of drug, what is the chance that his usage will go undetected?

Q3.3.24

A man has two lights in his well house to keep the pipes from freezing in winter. He checks the lights daily. Each light has probability 0.002 of burning out before it is checked the next day (independently of the other light).

a. If the lights are wired in parallel one will continue to shine even if the other burns out. In this situation, compute the probability that at least one light will continue to shine for the full 24 hours. Note the greatly increased reliability of the system of two bulbs over that of a single bulb.
b. If the lights are wired in series neither one will continue to shine even if only one of them burns out. In this situation, compute the probability that at least one light will continue to shine for the full 24 hours. Note the slightly decreased reliability of the system of two bulbs over that of a single bulb.

Q3.3.25

An accountant has observed that 5% of all copies of a particular two-part form have an error in Part I, and 2% have an error in Part II. If the errors occur independently, find the probability that a randomly selected form will be error-free.

Q3.3.26

A box contains 20 screws which are identical in size, but 12 of which are zinc coated and 8 of which are not. Two screws are selected at random, without replacement.

a. Find the probability that both are zinc coated.
b. Find the probability that at least one is zinc coated.

Additional Exercises

Q3.3.27

Events A and B are mutually exclusive. Find $P(A \mid B)$.

Q3.3.28

The city council of a particular city is composed of five members of party A, four members of party B, and three independents. Two council members are randomly selected to form an investigative committee.

a. Find the probability that both are from party A.
b. Find the probability that at least one is an independent.
c. Find the probability that the two have different party affiliations (that is, not both A, not both B, and not both independent).

Q3.3.29

A basketball player makes 60% of the free throws that he attempts, except that if he has just tried and missed a free throw then his chances of making a second one go down to only 30%. Suppose he has just been awarded two free throws.

a. Find the probability that he makes both.
b. Find the probability that he makes at least one. (A tree diagram could help.)

Q3.3.30

An economist wishes to ascertain the proportion p of the population of individual taxpayers who have purposely submitted fraudulent information on an income tax return. To truly guarantee anonymity of the taxpayers in a random survey, taxpayers questioned are given the following instructions.

a. Flip a coin.
b. If the coin lands heads, answer "Yes" to the question "Have you ever submitted fraudulent information on a tax return?" even if you have not.
c. If the coin lands tails, give a truthful "Yes" or "No" answer to the question "Have you ever submitted fraudulent information on a tax return?"

The questioner is not told how the coin landed, so he does not know if a "Yes" answer is the truth or is given only because of the coin toss.

a. Using the Probability Rule for Complements and the independence of the coin toss and the taxpayers' status fill in the empty cells in the two-way contingency table shown. Assume that the coin is fair. Each cell except the two in the bottom row will contain the unknown proportion (or probability) p.

Status	Coin		Probability
	H	T	
Fraud			p
No fraud			
Probability			1

b. The only information that the economist sees are the entries in the following table:

$Response$	$"Yes"$	$"No"$
$Proportion$	r	s

(5.E.9)

Equate the entry in the one cell in the table in (a) that corresponds to the answer "No" to the number s to obtain the formula that expresses the unknown number p in terms of the known number s.

c. Equate the sum of the entries in the three cells in the table in (a) that together correspond to the answer "Yes" to the number r to obtain the formula that expresses the unknown number p in terms of the known number r.

d. Use the fact that $r+s=1$ (since they are the probabilities of complementary events) to verify that the formulas in (b) and (c) give the same value for p. (For example, insert $s=1-r$ into the formula in (b) to obtain the formula in (c)).

e. Suppose a survey of $1,200$ taxpayers is conducted and 690 respond "Yes" (truthfully or not) to the question "Have you ever submitted fraudulent information on a tax return?" Use the answer to either (b) or (c) to estimate the true proportion p of all individual taxpayers who have purposely submitted fraudulent information on an income tax return.

Answers

1. a. 0.6
   b. 0.4
   c. not independent
2.
3. a. 0.22
   b. 0.81
   c. 0.27
4.
5. a. 0
   b. 0
6.
7. a. 0.5
   b. 0.4
   c. 0.6
8.
9. a. 0.25
   b. 0.33
   c. 0
   d. 0.25
10.
11. a. $P(A)=0.3,\ P(R)=0.4,\ P(A\cap R)=0.12$
    b. independent
    c. without computation 0.3
12.
13. a. Insufficient information. The events *A* and *B* are not known to be either independent or mutually exclusive.
    b. 0.21
    c. 0
14.
15. a. 0.25
    b. 0.02
16.
17. a. 0.5
    b. 0.57
    c. 0.875
    d. 0.75
18.

19. a. 0.4
    b. 0.43
    c. 0.38
20.
21. a. 0.15
    b. 0.14
    c. 0.25
    d. not independent
22.
23. 0.0375
24.
25. 0.931
26.
27. 0
28.
29. a. 0.36
    b. 0.72

Contributor

- Anonymous

CHAPTER OVERVIEW

6: Discrete Random Variables

It is often the case that a number is naturally associated to the outcome of a random experiment: the number of boys in a three-child family, the number of defective light bulbs in a case of 100 bulbs, the length of time until the next customer arrives at the drive-through window at a bank. Such a number varies from trial to trial of the corresponding experiment, and does so in a way that cannot be predicted with certainty; hence, it is called a random variable. In this chapter and the next we study such variables.

6.1: Random Variables

6.2: Probability Distributions for Discrete Random Variables

6.3: The Binomial Distribution

6.E: Discrete Random Variables (Exercises)

6.1: Random Variables

Learning Objectives

- To learn the concept of a random variable.
- To learn the distinction between discrete and continuous random variables.

Definition: random variable

A random variable is a numerical quantity that is generated by a random experiment.

We will denote random variables by capital letters, such as X or Z, and the actual values that they can take by lowercase letters, such as x and z.

Table 6.1.1 gives four examples of random variables. In the second example, the three dots indicates that every counting number is a possible value for X. Although it is highly unlikely, for example, that it would take 50 tosses of the coin to observe heads for the first time, nevertheless it is conceivable, hence the number 50 is a possible value. The set of possible values is infinite, but is still at least countable, in the sense that all possible values can be listed one after another. In the last two examples, by way of contrast, the possible values cannot be individually listed, but take up a whole interval of numbers. In the fourth example, since the light bulb could conceivably continue to shine indefinitely, there is no natural greatest value for its lifetime, so we simply place the symbol ∞ for infinity as the right endpoint of the interval of possible values.

Table 6.1.1: Four Random Variables

Experiment	Number X	Possible Values of X
Roll two fair dice	Sum of the number of dots on the top faces	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Flip a fair coin repeatedly	Number of tosses until the coin lands heads	1, 2, 3,4, …
Measure the voltage at an electrical outlet	Voltage measured	$118 \le x \le 122$
Operate a light bulb until it burns out	Time until the bulb burns out	$0 \le x < \infty$

Definition: discrete random variable

A random variable is called discrete if it has either a finite or a countable number of possible values. A random variable is called continuous if its possible values contain a whole interval of numbers.

The examples in the table are typical in that discrete random variables typically arise from a counting process, whereas continuous random variables typically arise from a measurement.

Key Takeaway

- A random variable is a number generated by a random experiment.
- A random variable is called discrete if its possible values form a finite or countable set.
- A random variable is called continuous if its possible values contain a whole interval of numbers.

6.2: Probability Distributions for Discrete Random Variables

Learning Objectives

- To learn the concept of the probability distribution of a discrete random variable.
- To learn the concepts of the mean, variance, and standard deviation of a discrete random variable, and how to compute them.

Associated to each possible value x of a discrete random variable X is the probability $P(x)$ that X will take the value x in one trial of the experiment.

Definition: probability distribution

The probability distribution of a discrete random variable X is a list of each possible value of X together with the probability that X takes that value in one trial of the experiment.

The probabilities in the probability distribution of a random variable X must satisfy the following two conditions:

- Each probability $P(x)$ must be between 0 and 1:

$$0 \le P(x) \le 1.$$

- The sum of all the possible probabilities is 1:

$$\sum P(x) = 1.$$

Example 6.2.1: two Fair Coins

A fair coin is tossed twice. Let X be the number of heads that are observed.

a. Construct the probability distribution of X.
b. Find the probability that at least one head is observed.

Solution

a. The possible values that X can take are 0, 1, and 2. Each of these numbers corresponds to an event in the sample space $S = \{hh, ht, th, tt\}$ of equally likely outcomes for this experiment:

$$X = 0 \text{ to } \{tt\},\ X = 1 \text{ to } \{ht, th\},\ \text{and } X = 2 \text{ to } hh.$$

The probability of each of these events, hence of the corresponding value of X, can be found simply by counting, to give

x	0	1	2
$P(x)$	0.25	0.50	0.25

This table is the probability distribution of X.

b. "At least one head" is the event $X \ge 1$, which is the union of the mutually exclusive events $X = 1$ and $X = 2$. Thus

$$P(X \ge 1) = P(1) + P(2) = 0.50 + 0.25$$
$$= 0.75$$

A histogram that graphically illustrates the probability distribution is given in Figure 6.2.1.

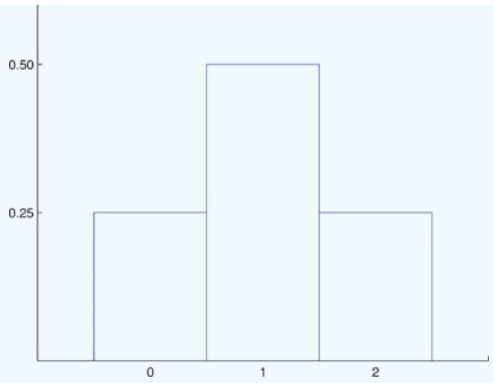


Figure 6.2.1: *Probability Distribution for Tossing a Fair Coin Twice*

Example 6.2.2: Two Fair Dice

A pair of fair dice is rolled. Let X denote the sum of the number of dots on the top faces.

a. Construct the probability distribution of X for a paid of fair dice.
b. Find $P(X \geq 9)$.
c. Find the probability that X takes an even value.

Solution

The sample space of equally likely outcomes is

$$\begin{array}{cccccc} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{array}$$

where the first digit is die 1 and the second number is die 2.

a. The possible values for X are the numbers 2 through 12. $X = 2$ is the event $\{11\}$, so $P(2) = 1/36$. $X = 3$ is the event $\{12, 21\}$, so $P(3) = 2/36$. Continuing this way we obtain the following table

x	2	3	4	5	6	7	8	9	10	11	12
$P(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

This table is the probability distribution of X.

b. The event $X \geq 9$ is the union of the mutually exclusive events $X = 9$, $X = 10$, $X = 11$, and $X = 12$. Thus

$$\begin{aligned} P(X \geq 9) &= P(9) + P(10) + P(11) + P(12) \\ &= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} \\ &= \frac{10}{36} \\ &= 0.2\overline{7} \end{aligned}$$

c. Before we immediately jump to the conclusion that the probability that X takes an even value must be 0.5, note that X takes six different even values but only five different odd values. We compute

$$
\begin{aligned}
P(X \text{ is even}) &= P(2)+P(4)+P(6)+P(8)+P(10)+P(12) \\
&= \frac{1}{36}+\frac{3}{36}+\frac{5}{36}+\frac{5}{36}+\frac{3}{36}+\frac{1}{36} \\
&= \frac{18}{36} \\
&= 0.5
\end{aligned}
$$

A histogram that graphically illustrates the probability distribution is given in Figure 6.2.2.

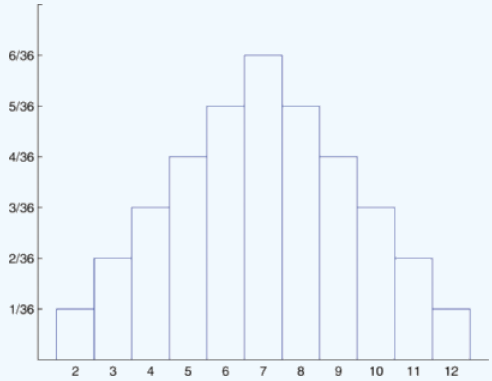


Figure 6.2.2: *Probability Distribution for Tossing Two Fair Dice*

The Mean and Standard Deviation of a Discrete Random Variable

Definition: mean

The *mean* (also called the "expectation value" or "expected value") of a discrete random variable X is the number

$$\mu = E(X) = \sum xP(x) \tag{6.2.1}$$

The mean of a random variable may be interpreted as the average of the values assumed by the random variable in repeated trials of the experiment.

Example 6.2.3

Find the mean of the discrete random variable X whose probability distribution is

x	-2	1	2	3.5
$P(x)$	0.21	0.34	0.24	0.21

Solution

Using the definition of mean (Equation 6.2.1) gives

$$
\begin{aligned}
\mu &= \sum xP(x) \\
&= (-2)(0.21)+(1)(0.34)+(2)(0.24)+(3.5)(0.21) \\
&= 1.135
\end{aligned}
$$

Example 6.2.4

A service organization in a large town organizes a raffle each month. One thousand raffle tickets are sold for $\$1$ each. Each has an equal chance of winning. First prize is $\$300$, second prize is $\$200$, and third prize is $\$100$. Let X denote the net gain from the purchase of one ticket.

a. Construct the probability distribution of X.
b. Find the probability of winning any money in the purchase of one ticket.

c. Find the expected value of X, and interpret its meaning.

Solution

a. If a ticket is selected as the first prize winner, the net gain to the purchaser is the $300 prize less the $1 that was paid for the ticket, hence $X = 300 - 11 = 299$. There is one such ticket, so $P(299) = 0.001$. Applying the same "income minus outgo" principle to the second and third prize winners and to the 997 losing tickets yields the probability distribution:

x	299	199	99	-1
$P(x)$	0.001	0.001	0.001	0.997

b. Let W denote the event that a ticket is selected to win one of the prizes. Using the table

$$\begin{aligned} P(W) &= P(299) + P(199) + P(99) = 0.001 + 0.001 + 0.001 \\ &= 0.003 \end{aligned}$$

c. Using the definition of expected value (Equation 6.2.1),

$$\begin{aligned} E(X) &= (299)\cdot(0.001) + (199)\cdot(0.001) + (99)\cdot(0.001) + (-1)\cdot(0.997) \\ &= -0.4 \end{aligned}$$

The negative value means that one loses money on the average. In particular, if someone were to buy tickets repeatedly, then although he would win now and then, on average he would lose 40 cents per ticket purchased.

The concept of expected value is also basic to the insurance industry, as the following simplified example illustrates.

Example 6.2.5

A life insurance company will sell a $200,000 one-year term life insurance policy to an individual in a particular risk group for a premium of $195. Find the expected value to the company of a single policy if a person in this risk group has a 99.97% chance of surviving one year.

Solution

Let X denote the net gain to the company from the sale of one such policy. There are two possibilities: the insured person lives the whole year or the insured person dies before the year is up. Applying the "income minus outgo" principle, in the former case the value of X is $195 - 0$; in the latter case it is $195 - 200,000 = -199,805$. Since the probability in the first case is 0.9997 and in the second case is $1 - 0.9997 = 0.0003$, the probability distribution for X is:

x	195	$-199,805$
$P(x)$	0.9997	0.0003

Therefore

$$\begin{aligned} E(X) &= \sum xP(x) \\ &= (195)\cdot(0.9997) + (-199,805)\cdot(0.0003) \\ &= 135 \end{aligned}$$

Occasionally (in fact, 3 times in 10,000) the company loses a large amount of money on a policy, but typically it gains $195, which by our computation of $E(X)$ works out to a net gain of $135 per policy sold, on average.

Definition: variance

The *variance* (σ^2) of a discrete random variable X is the number

$$\sigma^2 = \sum (x-\mu)^2 P(x) \tag{6.2.2}$$

which by algebra is equivalent to the formula

$$\sigma^2 = \left[\sum x^2 P(x)\right] - \mu^2 \tag{6.2.3}$$

Definition: standard deviation

The standard deviation, σ, *of a discrete random variable* X *is the square root of its variance, hence is given by the formulas*

$$\sigma = \sqrt{\sum (x-\mu)^2 P(x)} = \sqrt{\left[\sum x^2 P(x)\right] - \mu^2} \tag{6.2.4}$$

The variance and standard deviation of a discrete random variable X may be interpreted as measures of the variability of the values assumed by the random variable in repeated trials of the experiment. The units on the standard deviation match those of X.

Example 6.2.6

A discrete random variable X has the following probability distribution:

x	-1	0	1	4
$P(x)$	0.2	0.5	a	0.1

(6.2.5)

A histogram that graphically illustrates the probability distribution is given in Figure 6.2.3.

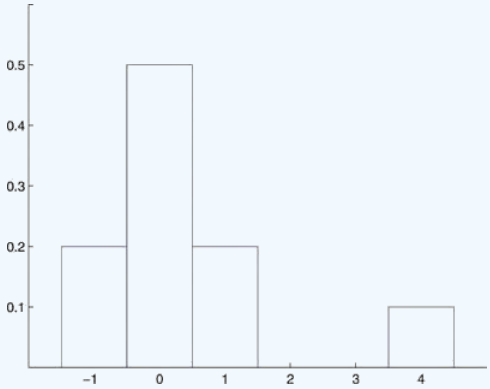


Figure 6.2.3: Probability Distribution of a Discrete Random Variable

Compute each of the following quantities.

a. a.
b. $P(0)$.
c. $P(X > 0)$.
d. $P(X \geq 0)$.
e. $P(X \leq -2)$.
f. The mean μ of X.
g. The variance σ^2 of X.
h. The standard deviation σ of X.

Solution

a. Since all probabilities must add up to 1,

$$a = 1 - (0.2 + 0.5 + 0.1) = 0.2$$

b. Directly from the table, P(0)=0.5

$$P(0) = 0.5$$

c. From Table 6.2.5,

$$P(X > 0) = P(1) + P(4) = 0.2 + 0.1 = 0.3$$

d. From Table 6.2.5,

$$P(X \geq 0) = P(0) + P(1) + P(4) = 0.5 + 0.2 + 0.1 = 0.8$$

e. Since none of the numbers listed as possible values for X is less than or equal to -2, the event $X \leq -2$ is impossible, so

$$P(X \leq -2) = 0$$

f. Using the formula in the definition of μ (Equation 6.2.1)

$$\begin{aligned} \mu &= \sum xP(x) \\ &= (-1) \cdot (0.2) + (0) \cdot (0.5) + (1) \cdot (0.2) + (4) \cdot (0.1) \\ &= 0.4 \end{aligned}$$

g. Using the formula in the definition of σ^2 (Equation 6.2.2) and the value of μ that was just computed,

$$\begin{aligned} \sigma^2 &= \sum (x - \mu)^2 P(x) \\ &= (-1 - 0.4)^2 \cdot (0.2) + (0 - 0.4)^2 \cdot (0.5) + (1 - 0.4)^2 \cdot (0.2) + (4 - 0.4)^2 \cdot (0.1) \\ &= 1.84 \end{aligned}$$

h. Using the result of part (g), $\sigma = \sqrt{1.84} = 1.3565$

Summary

- The probability distribution of a discrete random variable X is a listing of each possible value x taken by X along with the probability $P(x)$ that X takes that value in one trial of the experiment.
- The mean μ of a discrete random variable X is a number that indicates the average value of X over numerous trials of the experiment. It is computed using the formula $\mu = \sum xP(x)$.
- The variance σ^2 and standard deviation σ of a discrete random variable X are numbers that indicate the variability of X over numerous trials of the experiment. They may be computed using the formula $\sigma^2 = \left[\sum x^2 P(x)\right] - \mu^2$.

6.3: The Binomial Distribution

Learning Objectives

- To learn the concept of a binomial random variable.
- To learn how to recognize a random variable as being a binomial random variable.

The experiment of tossing a fair coin three times and the experiment of observing the genders according to birth order of the children in a randomly selected three-child family are completely different, but the random variables that count the number of heads in the coin toss and the number of boys in the family (assuming the two genders are equally likely) are the same random variable, the one with probability distribution

x	0	1	2	3
$P(x)$	0.125	0.375	0.375	0.125

A histogram that graphically illustrates this probability distribution is given in Figure 6.3.1. What is common to the two experiments is that we perform three identical and independent trials of the same action, each trial has only two outcomes (heads or tails, boy or girl), and the probability of success is the same number, 0.5, on every trial. The random variable that is generated is called the binomial random variable with parameters $n = 3$ and $p = 0.5$. This is just one case of a general situation.

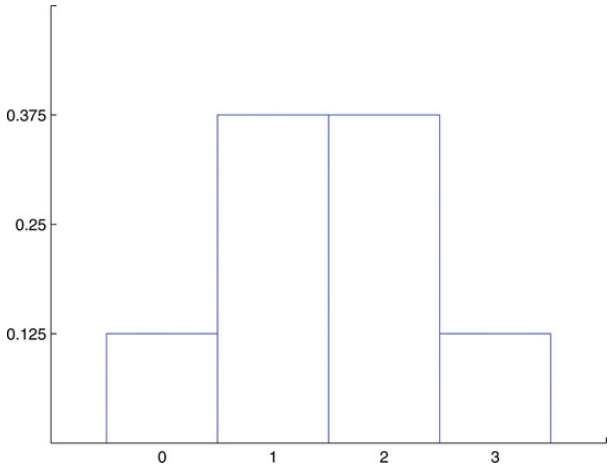


Figure 6.3.1: Probability Distribution for Three Coins and Three Children

Definition: binomial distribution

Suppose a random experiment has the following characteristics.

- There are n identical and independent trials of a common procedure.
- There are exactly two possible outcomes for each trial, one termed "success" and the other "failure."
- The probability of success on any one trial is the same number p.

Then the discrete random variable X that counts the number of successes in the n trials is the binomial random variable with parameters n and p. We also say that X has a binomial distribution with parameters n and p.

The following four examples illustrate the definition. Note how in every case "success" is the outcome that is counted, not the outcome that we prefer or think is better in some sense.

1. A random sample of 125 students is selected from a large college in which the proportion of students who are females is 57%. Suppose X denotes the number of female students in the sample. In this situation there are $n = 125$ identical and independent trials of a common procedure, selecting a student at random; there are exactly two possible outcomes for each trial, "success" (what we are counting, that the student be female) and "failure;" and finally the probability of success on any one trial is the same number $p = 0.57$. X is a binomial random variable with parameters $n = 125$ and $p = 0.57$.

2. A multiple-choice test has 15 questions, each of which has five choices. An unprepared student taking the test answers each of the questions completely randomly by choosing an arbitrary answer from the five provided. Suppose X denotes the number of answers that the student gets right. X is a binomial random variable with parameters $n = 15$ and $p = 1/5 = 0.20$.
3. In a survey of $1,000$ registered voters each voter is asked if he intends to vote for a candidate Titania Queen in the upcoming election. Suppose X denotes the number of voters in the survey who intend to vote for Titania Queen. X is a binomial random variable with $n = 1000$ and p equal to the true proportion of voters (surveyed or not) who intend to vote for Titania Queen.
4. An experimental medication was given to 30 patients with a certain medical condition. Suppose X denotes the number of patients who develop severe side effects. X is a binomial random variable with $n = 30$ and p equal to the true probability that a patient with the underlying condition will experience severe side effects if given that medication.

Probability Formula for a Binomial Random Variable

Often the most difficult aspect of working a problem that involves the binomial random variable is recognizing that the random variable in question has a binomial distribution. Once that is known, probabilities can be computed using the following formula.

If X is a binomial random variable with parameters n and p, then

$$P(x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}$$

where $q = 1 - p$ and where for any counting number m, $m!$ (read "m factorial") is defined by

$$0! = 1, 1! = 1, 2! = 1 \cdot 2, 3! = 1 \cdot 2 \cdot 3$$

and in general

$$m! = 1 \cdot 2 \cdots (m-1) \cdot m$$

Example 6.3.1

Seventeen percent of victims of financial fraud know the perpetrator of the fraud personally.

a. Use the formula to construct the probability distribution for the number X of people in a random sample of five victims of financial fraud who knew the perpetrator personally.
b. A investigator examines five cases of financial fraud every day. Find the most frequent number of cases each day in which the victim knew the perpetrator.
c. A investigator examines five cases of financial fraud every day. Find the average number of cases per day in which the victim knew the perpetrator.

Solution

The random variable X is binomial with parameters $n = 5$ and $p = 0.17$; $q = 1 - p = 0.83$. The possible values of X are $0, 1, 2, 3, 4,$ and 5

$$\begin{aligned} P(0) &= \frac{5!}{0!5!}(0.17)^0(0.83)^5 \\ &= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1)(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5)} 1 \cdot (0.3939040643) \\ &= 0.3939040643 \approx 0.3939 \end{aligned}$$

$$\begin{aligned} P(1) &= \frac{5!}{1!4!}(0.17)^1(0.83)^4 \\ &= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1)(1 \cdot 2 \cdot 3 \cdot 4)}(0.17) \cdot (0.47458321) \\ &= 5 \cdot (0.17) \cdot (0.47458321) \\ &= 0.4033957285 \approx 0.4034 \end{aligned}$$

$$
\begin{aligned}
P(2) &= \frac{5!}{2!3!}(0.17)^2(0.83)^3 \\
&= \frac{1\cdot 2\cdot 3\cdot 4\cdot 5}{(1\cdot 2)(1\cdot 2\cdot 3)}(0.0289)\cdot(0.571787) \\
&= 10\cdot(0.0289)\cdot(0.571787) \\
&= 0.165246443 \approx 0.1652
\end{aligned}
$$

The remaining three probabilities are computed similarly, to give the probability distribution

x	0	1	2	3	4	5
$P(x)$	0.3939	0.4034	0.1652	0.0338	0.0035	0.0001

The probabilities do not add up to exactly 1 because of rounding.

This probability distribution is represented by the histogram in Figure 6.3.2, which graphically illustrates just how improbable the events $X = 4$ and $X = 5$ are. The corresponding bar in the histogram above the number 4 is barely visible, if visible at all, and the bar above 5 is far too short to be visible.

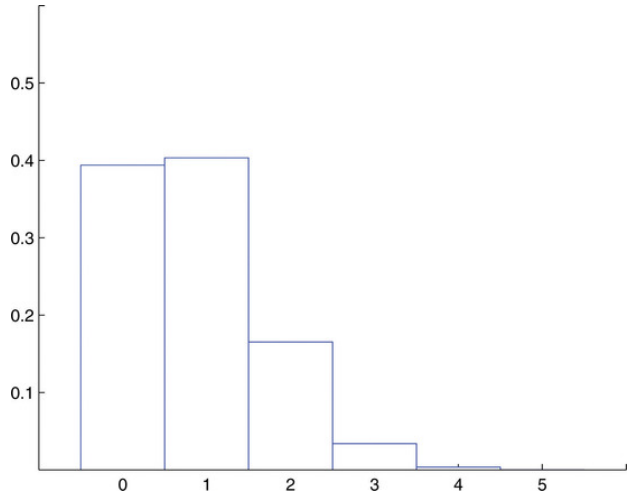


Figure 6.3.2: Probability Distribution of the Binomial Random Variable in Example 6.3.1

The value of X that is most likely is $X = 1$, so the most frequent number of cases seen each day in which the victim knew the perpetrator is one.

The average number of cases per day in which the victim knew the perpetrator is the mean of X, which is

$$
\begin{aligned}
\mu &= \sum xP(x) && (6.3.1)\\
&= 0\cdot 0.3939 + 1\cdot 0.4034 + 2\cdot 0.1652 + 3\cdot 0.0338 + 4\cdot 0.0035 + 5\cdot 0.0001 && (6.3.2)\\
&= 0.8497 && (6.3.3)
\end{aligned}
$$

Special Formulas for the Mean and Standard Deviation of a Binomial Random Variable

Since a binomial random variable is a discrete random variable, the formulas for its mean, variance, and standard deviation given in the previous section apply to it, as we just saw in Example 6.3.2 in the case of the mean. However, for the binomial random variable there are much simpler formulas.

If X is a binomial random variable with parameters n and p, then

$$\mu = np$$

$$\sigma^2 = npq$$

$$\sigma = \sqrt{npq}$$

where $q = 1 - p$.

Example 6.3.2

Find the mean and standard deviation of the random variable X of Example 6.3.1.

Solution

The random variable X is binomial with parameters $n = 5$ and $p = 0.17$, and $q = 1 - p = 0.83$. Thus its mean and standard deviation are

$$\mu = np = (5) \cdot (0.17) = 0.85 \quad \text{(exactly)}$$

and

$$\sigma = \sqrt{npq} = \sqrt{(5) \cdot (0.17) \cdot (0.83)} = \sqrt{0.7055} \approx 0.8399$$

The Cumulative Probability Distribution of a Binomial Random Variable

In order to allow a broader range of more realistic problems contains probability tables for binomial random variables for various choices of the parameters n and p. These tables are not the probability distributions that we have seen so far, but are **cumulative probability distributions**. In the place of the probability $P(x)$ the table contains the probability

$$P(X \leq x) = P(0) + P(1) + \ldots + P(x)$$

This is illustrated in Figure 6.3.3. The probability entered in the table corresponds to the area of the shaded region. The reason for providing a cumulative table is that in practical problems that involve a binomial random variable typically the probability that is sought is of the form $P(X \leq x)$ or $P(X \geq x)$. The cumulative table is much easier to use for computing $P(X \leq x)$ since all the individual probabilities have already been computed and added. The one table suffices for both $P(X \leq x)$ or $P(X \geq x)$ and can be used to readily obtain probabilities of the form $P(x)$, too, because of the following formulas. The first is just the Probability Rule for Complements.

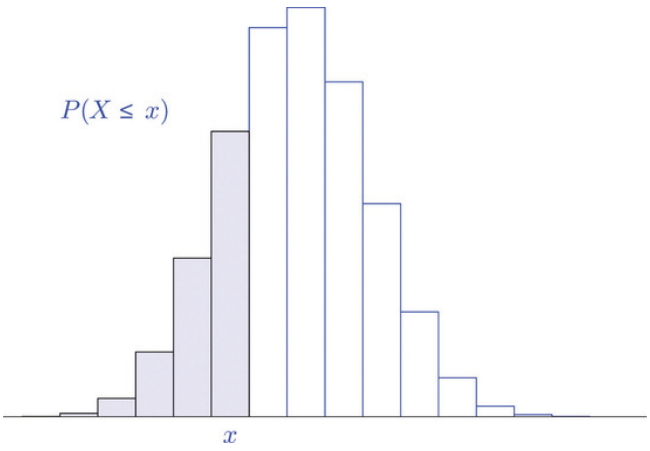


Figure 6.3.3: Cumulative Probabilities

If X is a discrete random variable, then

$$P(X \geq x) = 1 - P(X \leq x - 1)$$

and

$$P(x) = P(X \leq x) - P(X \leq x - 1)$$

Example 6.3.3

A student takes a ten-question true/false exam.

a. Find the probability that the student gets exactly six of the questions right simply by guessing the answer on every question.
b. Find the probability that the student will obtain a passing grade of 60% or greater simply by guessing.

Solution

Let X denote the number of questions that the student guesses correctly. Then X is a binomial random variable with parameters $n = 10$ and $p = 0.50$.

a. The probability sought is $P(6)$. The formula gives

$$P(6) = 10!(6!)(4!)(.5)6.54 = 0.205078125$$

Using the table,

$$P(6) = P(X \leq 6) - P(X \leq 5) = 0.8281 - 0.6230 = 0.2051$$

b. The student must guess correctly on at least 60% of the questions, which is $(0.60) \cdot (10) = 6$ questions. The probability sought is not $P(6)$ (an easy mistake to make), but

$$P(X \geq 6) = P(6) + P(7) + P(8) + P(9) + P(10)$$

Instead of computing each of these five numbers using the formula and adding them we can use the table to obtain

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.6230 = 0.3770$$

which is much less work and of sufficient accuracy for the situation at hand.

Example 6.3.4

An appliance repairman services five washing machines on site each day. One-third of the service calls require installation of a particular part.

a. The repairman has only one such part on his truck today. Find the probability that the one part will be enough today, that is, that at most one washing machine he services will require installation of this particular part.
b. Find the minimum number of such parts he should take with him each day in order that the probability that he have enough for the day's service calls is at least 95%.

Solution

Let X denote the number of service calls today on which the part is required. Then X is a binomial random variable with parameters $n = 5$ and $p = 1/3 = 0.\bar{3}$

a. Note that the probability in question is not $P(1)$, but rather $P(X \leq 1)$. Using the cumulative distribution table,

$$P(X \leq 1) = 0.4609$$

b. The answer is the smallest number x such that the table entry $P(X \leq x)$ is at least 0.9500. Since $P(X \leq 2) = 0.7901$ is less than 0.95, two parts are not enough. Since $P(X \leq 3) = 0.9547$ is as large as 0.95, three parts will suffice at least 95% of the time. Thus the minimum needed is three.

Summary

- The discrete random variable X that counts the number of successes in n identical, independent trials of a procedure that always results in either of two outcomes, “success” or “failure,” and in which the probability of success on each trial is the same number p, is called the binomial random variable with parameters n and p.
- There is a formula for the probability that the binomial random variable with parameters n and p will take a particular value x.
- There are special formulas for the mean, variance, and standard deviation of the binomial random variable with parameters n and p that are much simpler than the general formulas that apply to all discrete random variables.
- Cumulative probability distribution tables, when available, facilitate computation of probabilities encountered in typical practical situations.

6.E: Discrete Random Variables (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

4.1: Random Variables

Basic

1. Classify each random variable as either discrete or continuous.
   a. The number of arrivals at an emergency room between midnight and $6:00\ a.m.$
   b. The weight of a box of cereal labeled “18 ounces.”
   c. The duration of the next outgoing telephone call from a business office.
   d. The number of kernels of popcorn in a 1-pound container.
   e. The number of applicants for a job.
2. Classify each random variable as either discrete or continuous.
   a. The time between customers entering a checkout lane at a retail store.
   b. The weight of refuse on a truck arriving at a landfill.
   c. The number of passengers in a passenger vehicle on a highway at rush hour.
   d. The number of clerical errors on a medical chart.
   e. The number of accident-free days in one month at a factory.
3. Classify each random variable as either discrete or continuous.
   a. The number of boys in a randomly selected three-child family.
   b. The temperature of a cup of coffee served at a restaurant.
   c. The number of no-shows for every 100 reservations made with a commercial airline.
   d. The number of vehicles owned by a randomly selected household.
   e. The average amount spent on electricity each July by a randomly selected household in a certain state.
4. Classify each random variable as either discrete or continuous.
   a. The number of patrons arriving at a restaurant between $5:00\ p.m.$ and $6:00\ p.m.$
   b. The number of new cases of influenza in a particular county in a coming month.
   c. The air pressure of a tire on an automobile.
   d. The amount of rain recorded at an airport one day.
   e. The number of students who actually register for classes at a university next semester.
5. Identify the set of possible values for each random variable. (Make a reasonable estimate based on experience, where necessary.)
   a. The number of heads in two tosses of a coin.
   b. The average weight of newborn babies born in a particular county one month.
   c. The amount of liquid in a 12-ounce can of soft drink.
   d. The number of games in the next World Series (best of up to seven games).
   e. The number of coins that match when three coins are tossed at once.
6. Identify the set of possible values for each random variable. (Make a reasonable estimate based on experience, where necessary.)
   a. The number of hearts in a five-card hand drawn from a deck of 52 cards that contains 13 hearts in all.
   b. The number of pitches made by a starting pitcher in a major league baseball game.
   c. The number of breakdowns of city buses in a large city in one week.
   d. The distance a rental car rented on a daily rate is driven each day.
   e. The amount of rainfall at an airport next month.

Answers

1. a. discrete
   b. continuous
   c. continuous
   d. discrete

e. discrete

2.

3. a. discrete
   b. continuous
   c. discrete
   d. discrete
   e. continuous

4.

5. a. $\{0.1.2\}$
   b. an interval (a, b) (answers vary)
   c. an interval (a, b) (answers vary)
   d. $\{4, 5, 6, 7\}$
   e. $\{2, 3\}$

4.2: Probability Distributioins for Discrete Random Variables

Basic

1. Determine whether or not the table is a valid probability distribution of a discrete random variable. Explain fully.

   a.

x	-2	0	2	4
$P(x)$	0.3	0.5	0.2	0.1

   (6.E.1)

   b.

x	0.5	0.25	0.25
$P(x)$	-0.4	0.6	0.8

   (6.E.2)

   c.

x	1.1	2.5	4.1	4.6	5.3
$P(x)$	0.16	0.14	0.11	0.27	0.22

   (6.E.3)

2. Determine whether or not the table is a valid probability distribution of a discrete random variable. Explain fully.

   a.

x	0	1	2	3	4
$P(x)$	-0.25	0.50	0.35	0.10	0.30

   (6.E.4)

   b.

x	1	2	3
$P(x)$	0.325	0.406	0.164

   (6.E.5)

   c.

x	25	26	27	28	29
$P(x)$	0.13	0.27	0.28	0.18	0.14

   (6.E.6)

3. A discrete random variable X has the following probability distribution:

x	77	78	79	80	81
$P(x)$	0.15	0.15	0.20	0.40	0.10

   (6.E.7)

   Compute each of the following quantities.

   a. $P(80)$.
   b. $P(X > 80)$.
   c. $P(X \leq 80)$.
   d. The mean μ of X.
   e. The variance σ^2 of X.
   f. The standard deviation σ of X.

4. A discrete random variable X has the following probability distribution:

x	13	18	20	24	27
$P(x)$	0.22	0.25	0.20	0.17	0.16

(6.E.8)

Compute each of the following quantities.

a. $P(18)$.
b. $P(X > 18)$.
c. $P(X \leq 18)$.
d. The mean μ of X.
e. The variance σ^2 of X.
f. The standard deviation σ of X.

5. If each die in a pair is "loaded" so that one comes up half as often as it should, six comes up half again as often as it should, and the probabilities of the other faces are unaltered, then the probability distribution for the sum X of the number of dots on the top faces when the two are rolled is

x	2	3	4	5	6	7
$P(x)$	$\frac{1}{144}$	$\frac{4}{144}$	$\frac{8}{144}$	$\frac{12}{144}$	$\frac{16}{144}$	$\frac{22}{144}$

(6.E.9)

x	8	9	10	11	12
$P(x)$	$\frac{24}{144}$	$\frac{20}{144}$	$\frac{16}{144}$	$\frac{12}{144}$	$\frac{9}{144}$

(6.E.10)

Compute each of the following.

a. $P(5 \leq X \leq 9)$.
b. $P(X \geq 7)$.
c. The mean μ of X. (For fair dice this number is 7).
d. The standard deviation σ of X. (For fair dice this number is about 2.415).

Applications

6. Borachio works in an automotive tire factory. The number X of sound but blemished tires that he produces on a random day has the probability distribution

x	2	3	4	5
$P(x)$	0.48	0.36	0.12	0.04

(6.E.11)

a. Find the probability that Borachio will produce more than three blemished tires tomorrow.
b. Find the probability that Borachio will produce at most two blemished tires tomorrow.
c. Compute the mean and standard deviation of X. Interpret the mean in the context of the problem.

7. In a hamster breeder's experience the number X of live pups in a litter of a female not over twelve months in age who has not borne a litter in the past six weeks has the probability distribution

x	3	4	5	6	7	8	9
$P(x)$	0.04	0.10	0.26	0.31	0.22	0.05	0.02

(6.E.12)

a. Find the probability that the next litter will produce five to seven live pups.
b. Find the probability that the next litter will produce at least six live pups.
c. Compute the mean and standard deviation of X. Interpret the mean in the context of the problem.

8. The number X of days in the summer months that a construction crew cannot work because of the weather has the probability distribution

x	6	7	8	9	10
$P(x)$	0.03	0.08	0.15	0.20	0.19

(6.E.13)

x	11	12	13	14
$P(x)$	0.16	0.10	0.07	0.02

(6.E.14)

a. Find the probability that no more than ten days will be lost next summer.
b. Find the probability that from 8 to 12 days will be lost next summer.
c. Find the probability that no days at all will be lost next summer.
d. Compute the mean and standard deviation of X. Interpret the mean in the context of the problem.

9. Let X denote the number of boys in a randomly selected three-child family. Assuming that boys and girls are equally likely, construct the probability distribution of X.
10. Let X denote the number of times a fair coin lands heads in three tosses. Construct the probability distribution of X.
11. Five thousand lottery tickets are sold for $\$1$ each. One ticket will win $\$1,000$, two tickets will win $\$500$ each, and ten tickets will win $\$100$ each. Let X denote the net gain from the purchase of a randomly selected ticket.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X. Interpret its meaning.
    c. Compute the standard deviation σ of X.
12. Seven thousand lottery tickets are sold for $\$5$ each. One ticket will win $\$2,000$, two tickets will win $\$750$ each, and five tickets will win $\$100$ each. Let X denote the net gain from the purchase of a randomly selected ticket.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X. Interpret its meaning.
    c. Compute the standard deviation σ of X.
13. An insurance company will sell a $\$90,000$ one-year term life insurance policy to an individual in a particular risk group for a premium of $\$478$. Find the expected value to the company of a single policy if a person in this risk group has a 99.62% chance of surviving one year.
14. An insurance company will sell a $\$10,000$ one-year term life insurance policy to an individual in a particular risk group for a premium of $\$368$. Find the expected value to the company of a single policy if a person in this risk group has a 97.25% chance of surviving one year.
15. An insurance company estimates that the probability that an individual in a particular risk group will survive one year is 0.9825. Such a person wishes to buy a $\$150,000$ one-year term life insurance policy. Let C denote how much the insurance company charges such a person for such a policy.
    a. Construct the probability distribution of X. (Two entries in the table will contain C).
    b. Compute the expected value $E(X)$ of X.
    c. Determine the value C must have in order for the company to break even on all such policies (that is, to average a net gain of zero per policy on such policies).
    d. Determine the value C must have in order for the company to average a net gain of $\$250$ per policy on all such policies.
16. An insurance company estimates that the probability that an individual in a particular risk group will survive one year is 0.99. Such a person wishes to buy a $\$75,000$ one-year term life insurance policy. Let C denote how much the insurance company charges such a person for such a policy.
    a. Construct the probability distribution of X. (Two entries in the table will contain C).
    b. Compute the expected value $E(X)$ of X.
    c. Determine the value C must have in order for the company to break even on all such policies (that is, to average a net gain of zero per policy on such policies).
    d. Determine the value C must have in order for the company to average a net gain of $\$150$ per policy on all such policies.
17. A roulette wheel has 38 slots. Thirty-six slots are numbered from 1 to 36; half of them are red and half are black. The remaining two slots are numbered 0 and 00 and are green. In a $\$1$ bet on red, the bettor pays $\$1$ to play. If the ball lands in a red slot, he receives back the dollar he bet plus an additional dollar. If the ball does not land on red he loses his dollar. Let X denote the net gain to the bettor on one play of the game.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X, and interpret its meaning in the context of the problem.
    c. Compute the standard deviation of X.
18. A roulette wheel has 38 slots. Thirty-six slots are numbered from 1 to 36; the remaining two slots are numbered 0 and 00. Suppose the "number" 00 is considered not to be even, but the number 0 is still even. In a $\$1$ bet on even, the bettor pays $\$1$ to play. If the ball lands in an even numbered slot, he receives back the dollar he bet plus an additional dollar. If the ball does not land on an even numbered slot, he loses his dollar. Let X denote the net gain to the bettor on one play of the game.

a. Construct the probability distribution of X.
b. Compute the expected value $E(X)$ of X, and explain why this game is not offered in a casino (where 0 is not considered even).
c. Compute the standard deviation of X.

19. The time, to the nearest whole minute, that a city bus takes to go from one end of its route to the other has the probability distribution shown. As sometimes happens with probabilities computed as empirical relative frequencies, probabilities in the table add up only to a value other than 1.00 because of round-off error.

x	42	43	44	45	46	47
$P(x)$	0.10	0.23	0.34	0.25	0.05	0.02

(6.E.15)

a. Find the average time the bus takes to drive the length of its route.
b. Find the standard deviation of the length of time the bus takes to drive the length of its route.

20. Tybalt receives in the mail an offer to enter a national sweepstakes. The prizes and chances of winning are listed in the offer as: $\$5$ million, one chance in 65 million; $\$150,000$ one chance in 6.5 million; $\$5,000$ one chance in $650,000$; and $\$1,000$ one chance in $65,000$. If it costs Tybalt 44 cents to mail his entry, what is the expected value of the sweepstakes to him?

Additional Exercises

21. The number X of nails in a randomly selected 1-pound box has the probability distribution shown. Find the average number of nails per pound.

x	100	101	102
$P(x)$	0.01	0.96	0.03

(6.E.16)

22. Three fair dice are rolled at once. Let X denote the number of dice that land with the same number of dots on top as at least one other die. The probability distribution for X is

x	0	u	3
$P(x)$	p	$\frac{15}{36}$	$\frac{1}{36}$

(6.E.17)

a. Find the missing value u of X.
b. Find the missing probability p.
c. Compute the mean of X.
d. Compute the standard deviation of X.

23. Two fair dice are rolled at once. Let X denote the difference in the number of dots that appear on the top faces of the two dice. Thus for example if a one and a five are rolled, $X = 4$, and if two sixes are rolled, $X = 0$.
a. Construct the probability distribution for X.
b. Compute the mean μ of X.
c. Compute the standard deviation σ of X.

24. A fair coin is tossed repeatedly until either it lands heads or a total of five tosses have been made, whichever comes first. Let X denote the number of tosses made.
a. Construct the probability distribution for X.
b. Compute the mean μ of X.
c. Compute the standard deviation σ of X.

25. A manufacturer receives a certain component from a supplier in shipments of 100 units. Two units in each shipment are selected at random and tested. If either one of the units is defective the shipment is rejected. Suppose a shipment has 5 defective units.
a. Construct the probability distribution for the number X of defective units in such a sample. (A tree diagram is helpful).
b. Find the probability that such a shipment will be accepted.

26. Shylock enters a local branch bank at $4:30$ $p.m.$ every payday, at which time there are always two tellers on duty. The number X of customers in the bank who are either at a teller window or are waiting in a single line for the next available teller has the following probability distribution.

x	0	1	2	3
$P(x)$	0.135	0.192	0.284	0.230

(6.E.18)

x	4	5	6
$P(x)$	0.103	0.051	0.005

(6.E.19)

a. What number of customers does Shylock most often see in the bank the moment he enters?
b. What number of customers waiting in line does Shylock most often see the moment he enters?
c. What is the average number of customers who are waiting in line the moment Shylock enters?

27. The owner of a proposed outdoor theater must decide whether to include a cover that will allow shows to be performed in all weather conditions. Based on projected audience sizes and weather conditions, the probability distribution for the revenue X per night if the cover is not installed is

$Weather$	x	$P(x)$
$Clear$	$\$3,000$	0.61
$Threatening$	$\$2,800$	0.17
$LightRain$	$\$1,975$	0.11
$Show-cancelling\ rain$	$\$0$	0.11

(6.E.20)

The additional cost of the cover is $410,000. The owner will have it built if this cost can be recovered from the increased revenue the cover affords in the first ten 90-night seasons.

a. Compute the mean revenue per night if the cover is not installed.
b. Use the answer to (a) to compute the projected total revenue per 90-night season if the cover is not installed.
c. Compute the projected total revenue per season when the cover is in place. To do so assume that if the cover were in place the revenue each night of the season would be the same as the revenue on a clear night.
d. Using the answers to (b) and (c), decide whether or not the additional cost of the installation of the cover will be recovered from the increased revenue over the first ten years. Will the owner have the cover installed?

Answers

1. a. no: the sum of the probabilities exceeds 1
   b. no: a negative probability
   c. no: the sum of the probabilities is less than 1
2.
3. a. 0.4
   b. 0.1
   c. 0.9
   d. 79.15
   e. $\sigma^2 = 1.5275$
   f. $\sigma = 1.2359$
4.
5. a. 0.6528
   b. 0.7153
   c. $\mu = 7.8333$
   d. $\sigma^2 = 5.4866$
   e. $\sigma = 2.3424$
6.
7. a. 0.79
   b. 0.60
   c. $\mu = 5.8, \sigma = 1.2570$
8.
9.

x	0	1	2	3
$P(x)$	$1/8$	$3/8$	$3/8$	$1/8$

(6.E.21)

10.

11. a.

x	-1	999	499	99
$P(x)$	$\frac{4987}{5000}$	$\frac{1}{5000}$	$\frac{2}{5000}$	$\frac{10}{5000}$

(6.E.22)

b. -0.4
c. 17.8785

12.

13. 136

14.

15. a.

x	C	C	$-150,000$
$P(x)$	0.9825		0.0175

(6.E.23)

b. $C - 2625$
c. $C \geq 2625$
d. $C \geq 2875$

16.

17. a.

x	-1	1
$P(x)$	$\frac{20}{38}$	$\frac{18}{38}$

(6.E.24)

b. $E(X) = -0.0526$. In many bets the bettor sustains an average loss of about 5.25 cents per bet.
c. 0.9986

18.

19. a. 43.54
b. 1.2046

20.

21. 101.02

22.

23. a.

x	0	1	2	3	4	5
$P(x)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

(6.E.25)

b. 1.9444
c. 1.4326

24.

25. a.

x	0	1	2
$P(x)$	0.902	0.096	0.002

(6.E.26)

b. 0.902

26.

27. a. 2523.25
b. $227,092.5$
c. $270,000$
d. The owner will install the cover.

4.3: The Binomial Distribution

Basic

1. Determine whether or not the random variable X is a binomial random variable. If so, give the values of n and p. If not, explain why not.
   a. X is the number of dots on the top face of fair die that is rolled.
   b. X is the number of hearts in a five-card hand drawn (without replacement) from a well-shuffled ordinary deck.
   c. X is the number of defective parts in a sample of ten randomly selected parts coming from a manufacturing process in which 0.02% of all parts are defective.
   d. X is the number of times the number of dots on the top face of a fair die is even in six rolls of the die.
   e. X is the number of dice that show an even number of dots on the top face when six dice are rolled at once.
2. Determine whether or not the random variable X is a binomial random variable. If so, give the values of n and p. If not, explain why not.
   a. X is the number of black marbles in a sample of 5 marbles drawn randomly and without replacement from a box that contains 25 white marbles and 15 black marbles.
   b. X is the number of black marbles in a sample of 5 marbles drawn randomly and with replacement from a box that contains 25 white marbles and 15 black marbles.
   c. X is the number of voters in favor of proposed law in a sample $1,200$ randomly selected voters drawn from the entire electorate of a country in which 35% of the voters favor the law.
   d. X is the number of fish of a particular species, among the next ten landed by a commercial fishing boat, that are more than 13 inches in length, when 17% of all such fish exceed 13 inches in length.
   e. X is the number of coins that match at least one other coin when four coins are tossed at once.
3. X is a binomial random variable with parameters $n = 12$ and $p = 0.82$. Compute the probability indicated.
   a. $P(11)$
   b. $P(9)$
   c. $P(0)$
   d. $P(13)$
4. X is a binomial random variable with parameters $n = 16$ and $p = 0.74$. Compute the probability indicated.
   a. $P(14)$
   b. $P(4)$
   c. $P(0)$
   d. $P(20)$
5. X is a binomial random variable with parameters $n = 5, p = 0.5$. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $P(X \leq 3)$
   b. $P(X \geq 3)$
   c. $P(3)$
   d. $P(0)$
   e. $P(5)$
6. X is a binomial random variable with parameters $n = 5, p = 0.\overline{3}$. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $P(X \leq 2)$
   b. $P(X \geq 2)$
   c. $P(2)$
   d. $P(0)$
   e. $P(5)$
7. X is a binomial random variable with the parameters shown. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $n = 10, p = 0.25, P(X \leq 6)$
   b. $n = 10, p = 0.75, P(X \leq 6)$

c. $n = 15, p = 0.75, P(X \leq 6)$
d. $n = 15, p = 0.75, P(12)$
e. $n = 15, p = 0.\bar{6}, P(10 \leq X \leq 12)$

8. X is a binomial random variable with the parameters shown. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $n = 5, p = 0.05, P(X \leq 1)$
   b. $n = 5, p = 0.5, P(X \leq 1)$
   c. $n = 10, p = 0.75, P(X \leq 5)$
   d. $n = 10, p = 0.75, P(12)$
   e. $n = 10, p = 0.\bar{6}, P(5 \leq X \leq 8)$
9. X is a binomial random variable with the parameters shown. Use the special formulas to compute its mean μ and standard deviation σ.
   a. $n = 8, p = 0.43$
   b. $n = 47, p = 0.82$
   c. $n = 1200, p = 0.44$
   d. $n = 2100, p = 0.62$
10. X is a binomial random variable with the parameters shown. Use the special formulas to compute its mean μ and standard deviation σ.
    a. $n = 14, p = 0.55$
    b. $n = 83, p = 0.05$
    c. $n = 957, p = 0.35$
    d. $n = 1750, p = 0.79$
11. X is a binomial random variable with the parameters shown. Compute its mean μ and standard deviation σ in two ways, first using the tables in 7.1: Large Sample Estimation of a Population Mean in conjunction with the general formulas $\mu = \sum xP(x)$ and $\sigma = \sqrt{[\sum x^2 P(x)] - \mu^2}$, then using the special formulas $\mu = np$ and $\sigma = \sqrt{npq}$.
    a. $n = 5, p = 0.\bar{3}$
    b. $n = 10, p = 0.75$
12. X is a binomial random variable with the parameters shown. Compute its mean μ and standard deviation σ in two ways, first using the tables in 7.1: Large Sample Estimation of a Population Mean in conjunction with the general formulas $\mu = \sum xP(x)$ and $\sigma = \sqrt{[\sum x^2 P(x)] - \mu^2}$, then using the special formulas $\mu = np$ and $\sigma = \sqrt{npq}$.
    a. $n = 10, p = 0.25$
    b. $n = 15, p = 0.1$
13. X is a binomial random variable with parameters $n = 10$ and $p = 1/3$. Use the cumulative probability distribution for X that is given in 7.1: Large Sample Estimation of a Population Mean to construct the probability distribution of X.
14. X is a binomial random variable with parameters $n = 15$ and $p = 1/2$. Use the cumulative probability distribution for X that is given in 7.1: Large Sample Estimation of a Population Mean to construct the probability distribution of X.
15. In a certain board game a player's turn begins with three rolls of a pair of dice. If the player rolls doubles all three times there is a penalty. The probability of rolling doubles in a single roll of a pair of fair dice is $1/6$. Find the probability of rolling doubles all three times.
16. A coin is bent so that the probability that it lands heads up is $2/3$. The coin is tossed ten times.
    a. Find the probability that it lands heads up at most five times.
    b. Find the probability that it lands heads up more times than it lands tails up.

Applications

17. An English-speaking tourist visits a country in which 30% of the population speaks English. He needs to ask someone directions.
    a. Find the probability that the first person he encounters will be able to speak English.
    b. The tourist sees four local people standing at a bus stop. Find the probability that at least one of them will be able to speak English.
18. The probability that an egg in a retail package is cracked or broken is 0.025.

a. Find the probability that a carton of one dozen eggs contains no eggs that are either cracked or broken.
b. Find the probability that a carton of one dozen eggs has (i) at least one that is either cracked or broken; (ii) at least two that are cracked or broken.
c. Find the average number of cracked or broken eggs in one dozen cartons.

19. An appliance store sells 20 refrigerators each week. Ten percent of all purchasers of a refrigerator buy an extended warranty. Let X denote the number of the next 20 purchasers who do so.
    a. Verify that X satisfies the conditions for a binomial random variable, and find n and p.
    b. Find the probability that X is zero.
    c. Find the probability that X is two, three, or four.
    d. Find the probability that X is at least five.
20. Adverse growing conditions have caused 5% of grapefruit grown in a certain region to be of inferior quality. Grapefruit are sold by the dozen.
    a. Find the average number of inferior quality grapefruit per box of a dozen.
    b. A box that contains two or more grapefruit of inferior quality will cause a strong adverse customer reaction. Find the probability that a box of one dozen grapefruit will contain two or more grapefruit of inferior quality.
21. The probability that a 7-ounce skein of a discount worsted weight knitting yarn contains a knot is 0.25. Goneril buys ten skeins to crochet an afghan.
    a. Find the probability that (i) none of the ten skeins will contain a knot; (ii) at most one will.
    b. Find the expected number of skeins that contain knots.
    c. Find the most likely number of skeins that contain knots.
22. One-third of all patients who undergo a non-invasive but unpleasant medical test require a sedative. A laboratory performs 20 such tests daily. Let X denote the number of patients on any given day who require a sedative.
    a. Verify that X satisfies the conditions for a binomial random variable, and find n and p.
    b. Find the probability that on any given day between five and nine patients will require a sedative (include five and nine).
    c. Find the average number of patients each day who require a sedative.
    d. Using the cumulative probability distribution for X in 7.1: Large Sample Estimation of a Population Mean find the minimum number x_{min} of doses of the sedative that should be on hand at the start of the day so that there is a 99% chance that the laboratory will not run out.
23. About 2% of alumni give money upon receiving a solicitation from the college or university from which they graduated. Find the average number monetary gifts a college can expect from every $2,000$ solicitations it sends.
24. Of all college students who are eligible to give blood, about 18% do so on a regular basis. Each month a local blood bank sends an appeal to give blood to 250 randomly selected students. Find the average number of appeals in such mailings that are made to students who already give blood.
25. About 12% of all individuals write with their left hands. A class of 130 students meets in a classroom with 130 individual desks, exactly 14 of which are constructed for people who write with their left hands. Find the probability that exactly 14 of the students enrolled in the class write with their left hands.
26. A traveling salesman makes a sale on 65% of his calls on regular customers. He makes four sales calls each day.
    a. Construct the probability distribution of X, the number of sales made each day.
    b. Find the probability that, on a randomly selected day, the salesman will make a sale.
    c. Assuming that the salesman makes 20 sales calls per week, find the mean and standard deviation of the number of sales made *per week*.
27. A corporation has advertised heavily to try to insure that over half the adult population recognizes the brand name of its products. In a random sample of 20 adults, 14 recognized its brand name. What is the probability that 14 or more people in such a sample would recognize its brand name if the actual proportion p of all adults who recognize the brand name were only 0.50?

Additional Exercises

28. When dropped on a hard surface a thumbtack lands with its sharp point touching the surface with probability $2/3$; it lands with its sharp point directed up into the air with probability $1/3$. The tack is dropped and its landing position observed 15 times.
    a. Find the probability that it lands with its point in the air at least 7 times.

b. If the experiment of dropping the tack 15 times is done repeatedly, what is the average number of times it lands with its point in the air?

29. A professional proofreader has a 98% chance of detecting an error in a piece of written work (other than misspellings, double words, and similar errors that are machine detected). A work contains four errors.
    a. Find the probability that the proofreader will miss at least one of them.
    b. Show that two such proofreaders working independently have a 99.96% chance of detecting an error in a piece of written work.
    c. Find the probability that two such proofreaders working independently will miss at least one error in a work that contains four errors.
30. A multiple choice exam has 20 questions; there are four choices for each question.
    a. A student guesses the answer to every question. Find the chance that he guesses correctly between four and seven times.
    b. Find the minimum score the instructor can set so that the probability that a student will pass just by guessing is 20% or less.
31. In spite of the requirement that all dogs boarded in a kennel be inoculated, the chance that a healthy dog boarded in a clean, well-ventilated kennel will develop kennel cough from a carrier is 0.008.
    a. If a carrier (not known to be such, of course) is boarded with three other dogs, what is the probability that at least one of the three healthy dogs will develop kennel cough?
    b. If a carrier is boarded with four other dogs, what is the probability that at least one of the four healthy dogs will develop kennel cough?
    c. The pattern evident from parts (a) and (b) is that if $K+1$ dogs are boarded together, one a carrier and K healthy dogs, then the probability that at least one of the healthy dogs will develop kennel cough is $P(X \geq 1) = 1-(0.992)^K$, where X is the binomial random variable that counts the number of healthy dogs that develop the condition. Experiment with different values of K in this formula to find the maximum number $K+1$ of dogs that a kennel owner can board together so that if one of the dogs has the condition, the chance that another dog will be infected is less than 0.05.
32. Investigators need to determine which of 600 adults have a medical condition that affects 2% of the adult population. A blood sample is taken from each of the individuals.
    a. Show that the expected number of diseased individuals in the group of 600 is 12 individuals.
    b. Instead of testing all 600 blood samples to find the expected 12 diseased individuals, investigators group the samples into 60 groups of 10 each, mix a little of the blood from each of the 10 samples in each group, and test each of the 60 mixtures. Show that the probability that any such mixture will contain the blood of at least one diseased person, hence test positive, is about 0.18.
    c. Based on the result in (b), show that the expected number of mixtures that test positive is about 11. (Supposing that indeed 11 of the 60 mixtures test positive, then we know that none of the 490 persons whose blood was in the remaining 49 samples that tested negative has the disease. We have eliminated 490 persons from our search while performing only 60 tests.)

Answers

1. a. not binomial; not success/failure.
   b. not binomial; trials are not independent.
   c. binomial; $n = 10, p = 0.0002$
   d. binomial; $n = 6, p = 0.5$
   e. binomial; $n = 6, p = 0.5$
2.
3. a. 0.2434
   b. 0.2151
   c. $0.18^{12} \approx 0$
   d. 0
4.
5. a. 0.8125
   b. 0.5000
   c. 0.3125

d. 0.0313
e. 0.0312

6.

7. a. 0.9965
b. 0.2241
c. 0.0042
d. 0.2252
e. 0.5390

8.

9. a. $\mu = 3.44, \sigma = 1.4003$
b. $\mu = 38.54, \sigma = 2.6339$
c. $\mu = 528, \sigma = 17.1953$
d. $\mu = 1302, \sigma = 22.2432$

10.

11. a. $\mu = 1.6667, \sigma = 1.0541$
b. $\mu = 7.5, \sigma = 1.3693$

12.

13.

x	0	1	2	3
$P(x)$	0.0173	0.0867	0.1951	0.2602

(6.E.27)

x	4	5	6	7
$P(x)$	0.2276	0.1365	0.0569	0.0163

(6.E.28)

x	8	9	10
$P(x)$	0.0030	0.0004	0.0000

(6.E.29)

14.

15. 0.0046

16.

17. a. 0.3
b. 0.7599

18.

19. a. $n = 20, p = 0.1$
b. 0.1216
c. 0.5651
d. 0.0432

20.

21. a. 0.0563and 0.2440
b. 2.5
c. 2

22.

23. 40

24.

25. 0.1019

26.

27. 0.0577

28.

29. a. 0.0776
b. 0.9996
c. 0.0016

30.

31. a. 0.0238
    b. 0.0316
    c. 6

Contributor

- Anonymous

CHAPTER OVERVIEW

7: Continuous Random Variables

A random variable is called *continuous* if its set of possible values contains a whole interval of decimal numbers. In this chapter we investigate such random variables.

7.1: Continuous Random Variables
7.2: The Standard Normal Distribution
7.3: Probability Computations for General Normal Random Variables
7.4: Areas of Tails of Distributions
7.E: Continuous Random Variables (Exercises)

7.1: Continuous Random Variables

Learning Objectives

- To learn the concept of the probability distribution of a continuous random variable, and how it is used to compute probabilities.
- To learn basic facts about the family of normally distributed random variables.

The Probability Distribution of a Continuous Random Variable

For a discrete random variable X the probability that X assumes one of its possible values on a single trial of the experiment makes good sense. This is not the case for a continuous random variable. For example, suppose X denotes the length of time a commuter just arriving at a bus stop has to wait for the next bus. If buses run every 30 minutes without fail, then the set of possible values of X is the interval denoted $[0, 30]$, the set of all decimal numbers between 0 and 30. But although the number 7.211916 is a possible value of X, there is little or no meaning to the concept of the probability that the commuter will wait precisely 7.211916 minutes for the next bus. If anything the probability should be zero, since if we could meaningfully measure the waiting time to the nearest millionth of a minute it is practically inconceivable that we would ever get exactly 7.211916 minutes. More meaningful questions are those of the form: What is the probability that the commuter's waiting time is less than 10 minutes, or is between 5 and 10 minutes? In other words, with continuous random variables one is concerned not with the event that the variable assumes a single particular value, but with the event that the random variable assumes a value in a particular interval.

Definition: density function

The probability distribution of a continuous random variable X is an assignment of probabilities to intervals of decimal numbers using a function $f(x)$, called a density function, in the following way: the probability that X assumes a value in the interval $[a, b]$ is equal to the area of the region that is bounded above by the graph of the equation $y = f(x)$, bounded below by the x-axis, and bounded on the left and right by the vertical lines through a and b, as illustrated in Figure 7.1.1.

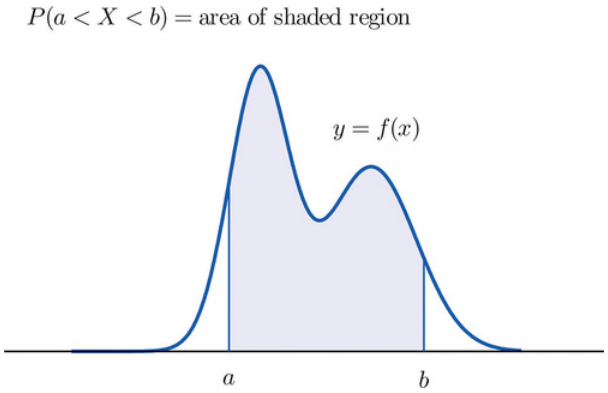


Figure 7.1.1: Probability Given as Area of a Region under a Curve

This definition can be understood as a natural outgrowth of the discussion in Section 2.1.3. There we saw that if we have in view a population (or a very large sample) and make measurements with greater and greater precision, then as the bars in the relative frequency histogram become exceedingly fine their vertical sides merge and disappear, and what is left is just the curve formed by their tops, as shown in Figure 2.1.5. Moreover the total area under the curve is 1, and the proportion of the population with measurements between two numbers a and b is the area under the curve and between a and b, as shown in Figure 2.1.6. If we think of X as a measurement to infinite precision arising from the selection of any one member of the population at random, then $P(a < X < b)$ is simply the proportion of the population with measurements between a and b, the curve in the relative frequency histogram is the density function for X, and we arrive at the definition just above.

- Every density function $f(x)$ must satisfy the following two conditions:
- For all numbers x, $f(x) \geq 0$, so that the graph of $y = f(x)$ never drops below the x-axis.
- The area of the region under the graph of $y = f(x)$ and above the x-axis is 1.

Because the area of a line segment is 0, the definition of the probability distribution of a continuous random variable implies that for any particular decimal number, say a, the probability that X assumes the exact value a is 0. This property implies that whether or not the endpoints of an interval are included makes no difference concerning the probability of the interval.

For any continuous random variable X:

$$P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$$

Example 7.1.1

A random variable X has the uniform distribution on the interval $[0, 1]$: the density function is $f(x) = 1$ if x is between 0 and 1 and $f(x) = 0$ for all other values of x, as shown in Figure 7.1.2.

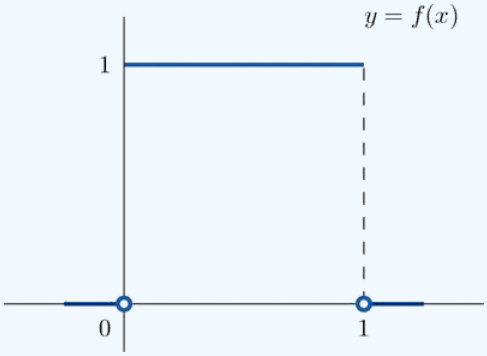


Figure 7.1.2: Uniform Distribution on [0,1].

1. Find $P(X > 0.75)$, the probability that X assumes a value greater than 0.75.
2. Find $P(X \le 0.2)$, the probability that X assumes a value less than or equal to 0.2.
3. Find $P(0.4 < X < 0.7)$, the probability that X assumes a value between 0.4 and 0.7.

Solution

1. $P(X > 0.75)$ is the area of the rectangle of height 1 and base length $1 - 0.75 = 0.25$, hence is $base \times height = (0.25) \cdot (1) = 0.25$. See Figure 7.1.3a.
2. $P(X \le 0.2)$ is the area of the rectangle of height 1 and base length $0.2 - 0 = 0.2$, hence is $base \times height = (0.2) \cdot (1) = 0.2$. See Figure 7.1.3b.
3. $P(0.4 < X < 0.7)$ is the area of the rectangle of height 1 and length $0.7 - 0.4 = 0.3$, hence is $base \times height = (0.3) \cdot (1) = 0.3$. See Figure 7.1.3c.

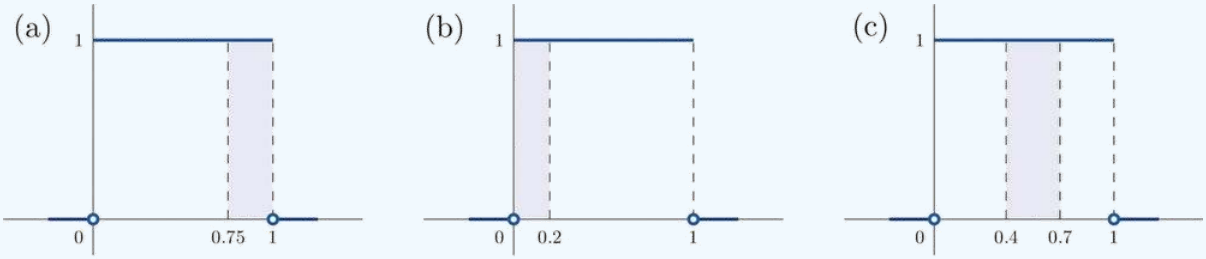


Figure 7.1.3: Probabilities from the Uniform Distribution on [0,1]

Example 7.1.2

A man arrives at a bus stop at a random time (that is, with no regard for the scheduled service) to catch the next bus. Buses run every 30 minutes without fail, hence the next bus will come any time during the next 30 minutes with evenly distributed probability (a uniform distribution). Find the probability that a bus will come within the next 10 minutes.

Solution

The graph of the density function is a horizontal line above the interval from 0 to 30 and is the x-axis everywhere else. Since the total area under the curve must be 1, the height of the horizontal line is $1/30$ (Figure 7.1.4). The probability sought is $P(0 \le X \le 10)$ .By definition, this probability is the area of the rectangular region bounded above by the horizontal line

$f(x) = 1/30$, bounded below by the x-axis, bounded on the left by the vertical line at 0 (the y-axis), and bounded on the right by the vertical line at 10. This is the shaded region in Figure 7.1.4. Its area is the base of the rectangle times its height, $(10) \cdot (1/30) = 1/3$. Thus $P(0 \le X \le 10) = 1/3$.

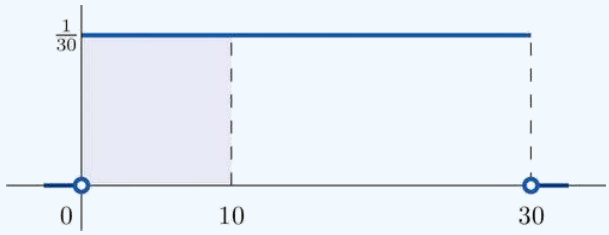


Figure 7.1.4: Probability of Waiting At Most 10 Minutes for a Bus

Normal Distributions

Most people have heard of the "bell curve." It is the graph of a specific density function $f(x)$ that describes the behavior of continuous random variables as different as the heights of human beings, the amount of a product in a container that was filled by a high-speed packing machine, or the velocities of molecules in a gas. The formula for $f(x)$ contains two parameters μ and σ that can be assigned any specific numerical values, so long as σ is positive. We will not need to know the formula for $f(x)$, but for those who are interested it is

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

where $\pi \approx 3.14159$ and $e \approx 2.71828$ is the base of the natural logarithms.

Each different choice of specific numerical values for the pair μ and σ gives a different bell curve. The value of μ determines the location of the curve, as shown in Figure 7.1.5. In each case the curve is symmetric about μ.

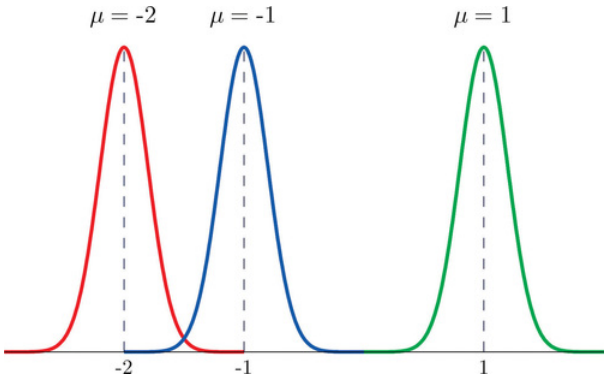


Figure 7.1.5: Bell Curves with σ = 0.25 and Different Values of μ

The value of σ determines whether the bell curve is tall and thin or short and squat, subject always to the condition that the total area under the curve be equal to 1. This is shown in Figure 7.1.6, where we have arbitrarily chosen to center the curves at $\mu = 6$.

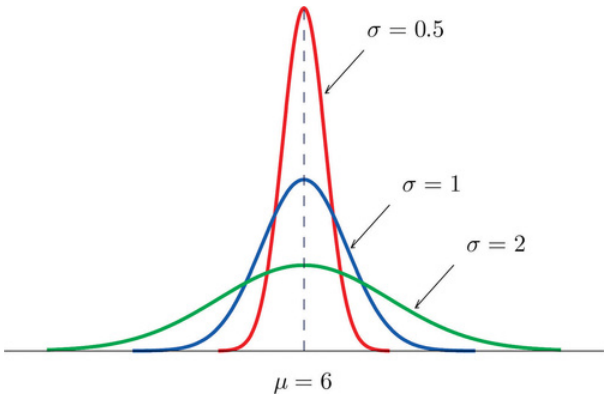


Figure 7.1.6: Bell Curves with $\mu = 6$ and Different Values of σ.

Definition: normal distribution

The probability distribution corresponding to the density function for the bell curve with parameters μ and σ is called the normal distribution with mean μ and standard deviation σ.

Definition: normally distributed random variable

A continuous random variable whose probabilities are described by the normal distribution with mean μ and standard deviation σ is called a normally distributed random variable, or a normal random variable for short, with mean μ and standard deviation σ.

Figure 7.1.7 shows the density function that determines the normal distribution with mean μ and standard deviation σ. We repeat an important fact about this curve: **The density curve for the normal distribution is symmetric about the mean.**

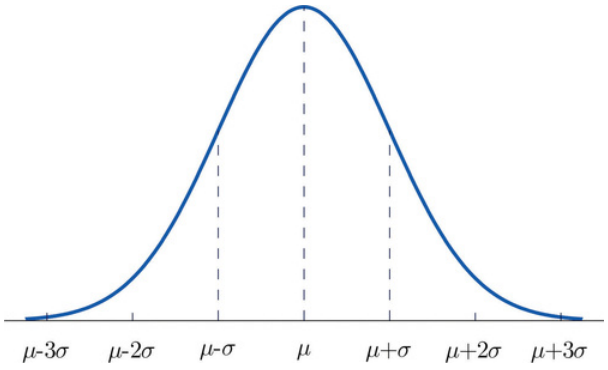


Figure 7.1.7: Density Function for a Normally Distributed Random Variable with Mean μ and Standard Deviation σ

Example 7.1.3

Heights of 25-year-old men in a certain region have mean 69.75 inches and standard deviation 2.59 inches. These heights are approximately normally distributed. Thus the height X of a randomly selected 25-year-old man is a normal random variable with mean $\mu = 69.75$ and standard deviation $\sigma = 2.59$. Sketch a qualitatively accurate graph of the density function for X. Find the probability that a randomly selected 25-year-old man is more than 69.75 inches tall.

Solution

The distribution of heights looks like the bell curve in Figure 7.1.8. The important point is that it is centered at its mean, 69.75, and is symmetric about the mean.

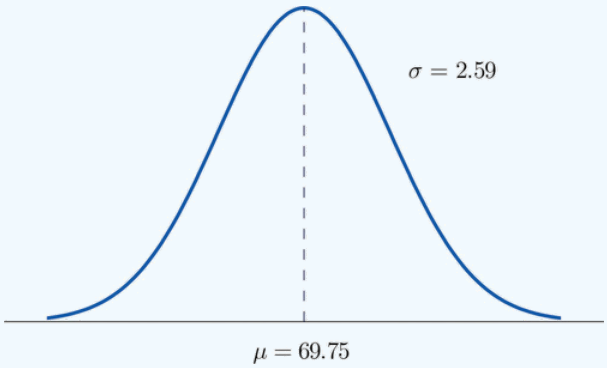


Figure 7.1.8: Density Function for Heights of 25-Year-Old Men

Since the total area under the curve is 1, by symmetry the area to the right of 69.75 is half the total, or 0.5. But this area is precisely the probability $P(X > 69.75)$, the probability that a randomly selected 25-year-old man is more than 69.75 inches tall. We will learn how to compute other probabilities in the next two sections.

Key Takeaway

- For a continuous random variable X the only probabilities that are computed are those of X taking a value in a specified interval.
- The probability that X take a value in a particular interval is the same whether or not the endpoints of the interval are included.
- The probability $P(a < X < b)$, that X take a value in the interval from a to b, is the area of the region between the vertical lines through a and b, above the x-axis, and below the graph of a function $f(x)$ called the density function.
- A normally distributed random variable is one whose density function is a bell curve.
- Every bell curve is symmetric about its mean and lies everywhere above the x-axis, which it approaches asymptotically (arbitrarily closely without touching).

7.2: The Standard Normal Distribution

Learning Objectives

- To learn what a standard normal random variable is.
- To learn how to compute probabilities related to a standard normal random variable.

Definition: standard normal random variable

A *standard normal random variable* is a normally distributed random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$. It will always be denoted by the letter Z.

The density function for a standard normal random variable is shown in Figure 7.2.1.

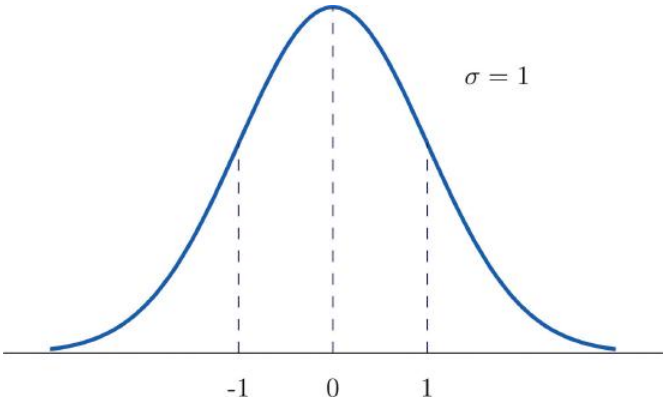


Figure 7.2.1: Density Curve for a Standard Normal Random Variable

To compute probabilities for Z we will not work with its density function directly but instead read probabilities out of Figure 7.2.2. The tables are tables of *cumulative* probabilities; their entries are probabilities of the form $P(Z < z)$. The use of the tables will be explained by the following series of examples.

Example 7.2.1

Find the probabilities indicated, where as always Z denotes a standard normal random variable.

1. $P(Z < 1.48)$.
2. $P(Z < -0.25)$.

Solution

1. Figure 7.2.3 shows how this probability is read directly from the table without any computation required. The digits in the ones and tenths places of 1.48, namely 1.4, are used to select the appropriate row of the table; the hundredths part of 1.48, namely 0.08, is used to select the appropriate column of the table. The four decimal place number in the interior of the table that lies in the intersection of the row and column selected, 0.9306, is the probability sought:

$$P(Z < 1.48) = 0.9306$$

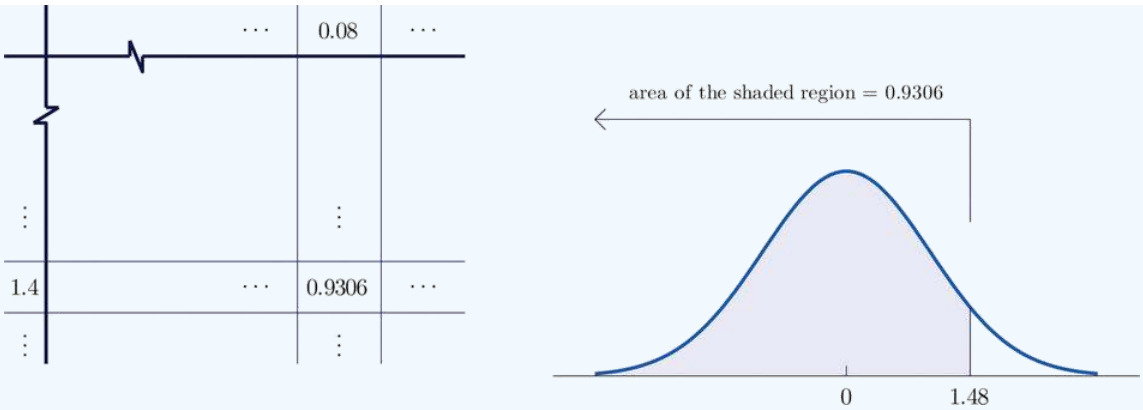


Figure 7.2.3: *Computing Probabilities Using the Cumulative Table*

2. The minus sign in -0.25 makes no difference in the procedure; the table is used in exactly the same way as in part (a): the probability sought is the number that is in the intersection of the row with heading -0.2 and the column with heading 0.05, the number 0.4013. Thus $P(Z < -0.25) = 0.4013$.

Example 7.2.2

Find the probabilities indicated.

1. $P(Z > 1.60)$.
2. $P(Z > -1.02)$.

Solution

1. Because the events $Z > 1.60$ and $Z \leq 1.60$ are complements, the *Probability Rule for Complements* implies that

$$P(Z > 1.60) = 1 - P(Z \leq 1.60)$$

Since inclusion of the endpoint makes no difference for the continuous random variable Z, $P(Z \leq 1.60) = P(Z < 1.60)$, which we know how to find from the table in Figure 7.2.2. The number in the row with heading 1.6 and in the column with heading 0.00 is 0.9452. Thus $P(Z < 1.60) = 0.9452$ so

$$P(Z > 1.60) = 1 - P(Z \leq 1.60) = 1 - 0.9452 = 0.0548$$

Figure 7.2.4 illustrates the ideas geometrically. Since the total area under the curve is 1 and the area of the region to the left of 1.60 is (from the table) 0.9452, the area of the region to the right of 1.60 must be $1 - 0.9452 = 0.0548$.

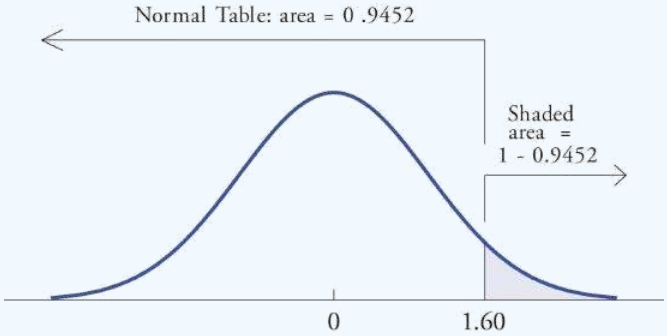


Figure 7.2.4: Computing a Probability for a Right Half-Line

2. The minus sign in -1.02 makes no difference in the procedure; the table is used in exactly the same way as in part (a). The number in the intersection of the row with heading -1.0 and the column with heading 0.02 is 0.1539. This means that $P(Z < -1.02) = P(Z \leq -1.02) = 0.1539$. Hence

$$P(Z > -1.02) = P(Z \leq -1.02) = 1 - 0.1539 = 0.8461$$

Example 7.2.3

Find the probabilities indicated.

1. $P(0.5 < Z < 1.57)$.
2. $P(-2.55 < Z < 0.09)$.

Solution

1. Figure 7.2.5 illustrates the ideas involved for intervals of this type. First look up the areas in the table that correspond to the numbers 0.5 (which we think of as 0.50 to use the table) and 1.57. We obtain 0.6915 and 0.9418, respectively. From the figure it is apparent that we must take the difference of these two numbers to obtain the probability desired. In symbols,

$$P(0.5 < Z < 1.57) = P(Z < 1.57) - P(Z < 0.50) = 0.9418 - 0.6915 = 0.2503$$

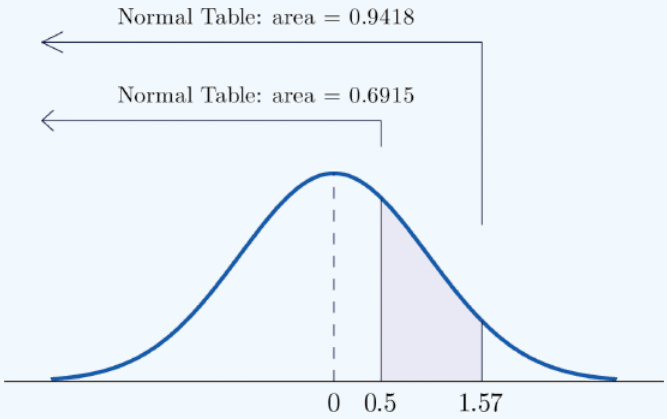


Figure 7.2.5: Computing a Probability for an Interval of Finite Length

2. The procedure for finding the probability that Z takes a value in a finite interval whose endpoints have opposite signs is exactly the same procedure used in part (a), and is illustrated in Figure 7.2.6 "Computing a Probability for an Interval of Finite Length". In symbols the computation is

$$P(-2.55 < Z < 0.09) = P(Z < 0.09) - P(Z < -2.55) = 0.5359 - 0.0054 = 0.5305$$

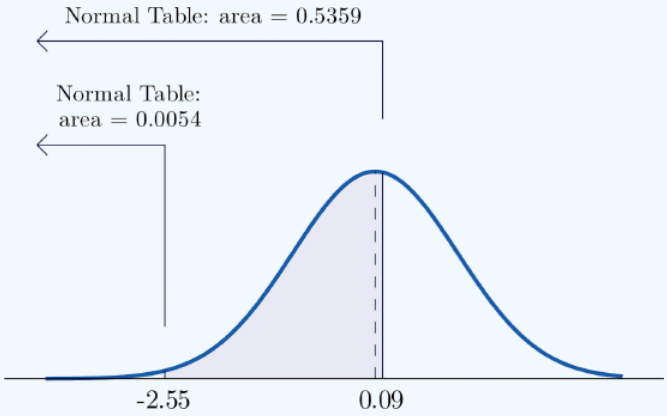


Figure 7.2.6: Computing a Probability for an Interval of Finite Length

The next example shows what to do if the value of Z that we want to look up in the table is not present there.

Example 7.2.4

Find the probabilities indicated.

1. $P(1.13 < Z < 4.16)$.
2. $P(-5.22 < Z < 2.15)$.

Solution

1. We attempt to compute the probability exactly as in Example 7.2.3 by looking up the numbers 1.13 and 4.16 in the table. We obtain the value 0.8708 for the area of the region under the density curve to left of 1.13 without any problem, but when we go to look up the number 4.16 in the table, it is not there. We can see from the last row of numbers in the table that the area to the left of 4.16 must be so close to 1 that to four decimal places it rounds to 1.0000. Therefore

$$P(1.13 < Z < 4.16) = 1.0000 - 0.8708 = 0.1292$$

2. Similarly, here we can read directly from the table that the area under the density curve and to the left of 2.15 is 0.9842, but −5.22 is too far to the left on the number line to be in the table. We can see from the first line of the table that the area to the left of −5.22 must be so close to 0 that to four decimal places it rounds to 0.0000. Therefore

$$P(-5.22 < Z < 2.15) = 0.9842 - 0.0000 = 0.9842$$

The final example of this section explains the origin of the proportions given in the Empirical Rule.

Example 7.2.5

Find the probabilities indicated.

1. $P(-1 < Z < 1)$.
2. $P(-2 < Z < 2)$.
3. $P(-3 < Z < 3)$.

Solution

1. Using the table as was done in Example 7.2.3 we obtain

$$P(-1 < Z < 1) = 0.8413 - 0.1587 = 0.6826$$

Since Z has mean 0 and standard deviation 1, for Z to take a value between −1 and 1 means that Z takes a value that is within one standard deviation of the mean. Our computation shows that the probability that this happens is about 0.68, the proportion given by the Empirical Rule for histograms that are mound shaped and symmetrical, like the bell curve.

2. Using the table in the same way,

$$P(-2 < Z < 2) = 0.9772 - 0.0228 = 0.9544$$

This corresponds to the proportion 0.95 for data within two standard deviations of the mean.

3. Similarly,

$$P(-3 < Z < 3) = 0.9987 - 0.0013 = 0.9974$$

which corresponds to the proportion 0.997 for data within three standard deviations of the mean.

Key takeaway

- A standard normal random variable Z is a normally distributed random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$.
- Probabilities for a standard normal random variable are computed using Figure 7.2.2.

7.3: Probability Computations for General Normal Random Variables

Learning Objectives

- To learn how to compute probabilities related to any normal random variable.

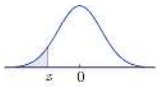
If X is any normally distributed normal random variable then Figure $7.3.1$ can also be used to compute a probability of the form $P(a < X < b)$ by means of the following equality.

equality

If X is a normally distributed random variable with mean μ and standard deviation σ, then

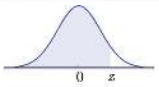
$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

where Z denotes a standard normal random variable. a can be any decimal number or $-\infty$; b can be any decimal number or ∞.



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8304	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981

2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure 7.3.1: Cumulative Normal Probability

The new endpoints $\frac{(a-\mu)}{\sigma}$ and $\frac{(b-\mu)}{\sigma}$ are the z-scores of a and b as defined in Chapter 2.

Figure 7.3.2 illustrates the meaning of the equality geometrically: the two shaded regions, one under the density curve for X and the other under the density curve for Z, have the same area. Instead of drawing both bell curves, though, we will always draw a single generic bell-shaped curve with both an x-axis and a z-axis below it.

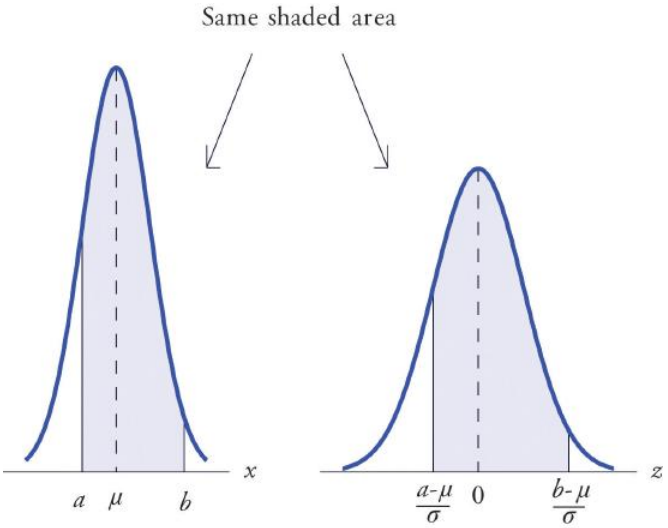


Figure 7.3.2: Probability for an Interval of Finite Length

Example 7.3.1

Let X be a normal random variable with mean $\mu = 10$ and standard deviation $\sigma = 2.5$. Compute the following probabilities.

1. $P(X < 14)$.
2. $P(8 < X < 14)$.

Solution

1. See Figure 7.3.3 "Probability Computation for a General Normal Random Variable".

$$\begin{aligned} P(X < 14) &= P\left(Z < \frac{14-\mu}{\sigma}\right) \\ &= P\left(Z < \frac{14-10}{2.5}\right) \\ &= P(Z < 1.60) \\ &= 0.9452 \end{aligned}$$

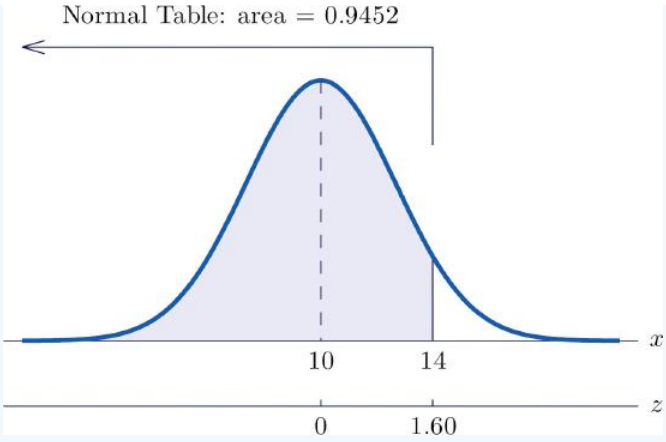


Figure 7.3.3: *Probability Computation for a General Normal Random Variable*

2. See Figure 7.3.4 "Probability Computation for a General Normal Random Variable".

$$\begin{aligned} P(8 < X < 14) &= P\left(\frac{8-10}{2.5} < Z < \frac{14-10}{2.5}\right) \\ &= P\left(-0.80 < Z < 1.60\right) \\ &= 0.9452 - 0.2119 \\ &= 0.7333 \end{aligned}$$

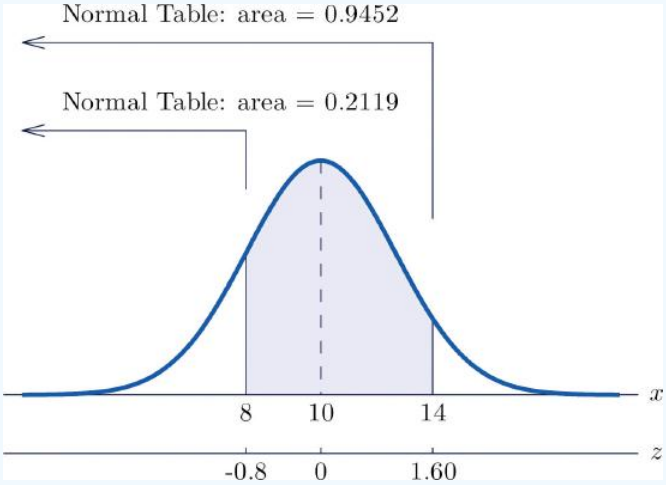


Figure 7.3.4: *Probability Computation for a General Normal Random Variable*

Example 7.3.2

The lifetimes of the tread of a certain automobile tire are normally distributed with mean $37,500$ miles and standard deviation $4,500$ miles. Find the probability that the tread life of a randomly selected tire will be between $30,000$ and $40,000$ miles.

Solution

Let X denote the tread life of a randomly selected tire. To make the numbers easier to work with we will choose thousands of miles as the units. Thus $\mu = 37.5$, $\sigma = 4.5$, and the problem is to compute $P(30 < X < 40)$. Figure 7.3.5 "Probability Computation for Tire Tread Wear" illustrates the following computation:

$$
\begin{aligned}
P(30 < X < 40) &= P\left(\frac{30-\mu}{\sigma} < Z < \frac{40-\mu}{\sigma}\right) \\
&= P\left(\frac{30-37.5}{4.5} < Z < \frac{40-37.5}{4.5}\right) \\
&= P\left(-1.67 < Z < 0.56\right) \\
&= 0.7123 - 0.0475 \\
&= 0.6648
\end{aligned}
$$

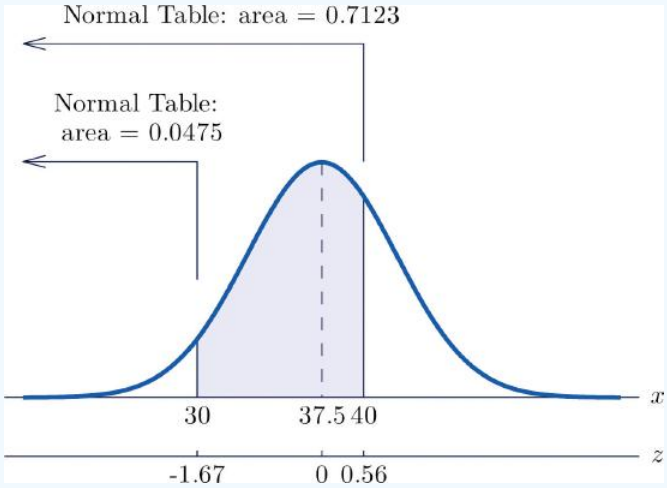


Figure 7.3.5: Probability Computation for Tire Tread Wear

Note that the two z-scores were rounded to two decimal places in order to use Figure 7.3.1 "Cumulative Normal Probability".

Example 7.3.3

Scores on a standardized college entrance examination (*CEE*) are normally distributed with mean 510 and standard deviation 60. A selective university considers for admission only applicants with *CEE* scores over 650. Find percentage of all individuals who took the *CEE* who meet the university's *CEE* requirement for consideration for admission.

Solution

Let X denote the score made on the *CEE* by a randomly selected individual. Then X is normally distributed with mean 510 and standard deviation 60. The probability that X lie in a particular interval is the same as the proportion of all exam scores that lie in that interval. Thus the solution to the problem is $P(X > 650)$, expressed as a percentage. Figure 7.3.6 "Probability Computation for Exam Scores" illustrates the following computation:

$$
\begin{aligned}
P(X > 650) &= P\left(Z > \frac{650-\mu}{\sigma}\right) \\
&= P\left(Z > \frac{650-510}{60}\right) \\
&= P(Z > 2.33) \\
&= 1 - 0.9901 \\
&= 0.0099
\end{aligned}
$$

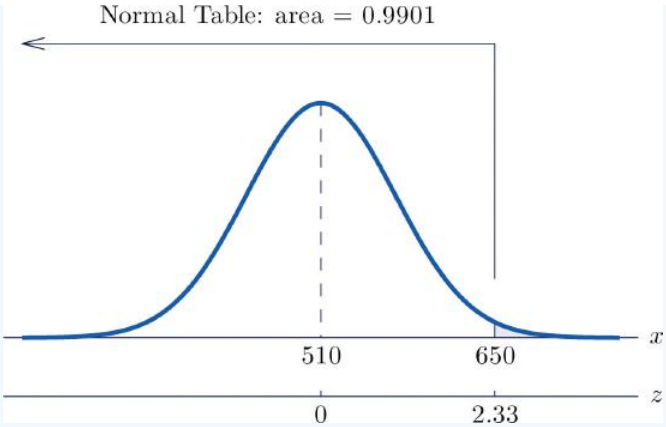


Figure 7.3.6: *Probability Computation for Exam Scores*

The proportion of all *CEE* scores that exceed 650 is 0.0099, hence 0.99% or about 1% do.

- Probabilities for a general normal random variable are computed using Figure $7.3.1$ after converting x-values to z-scores.

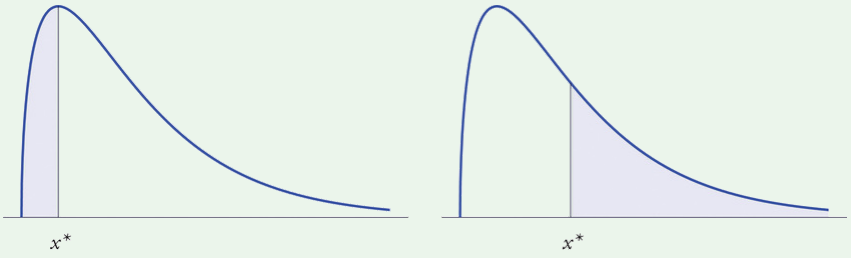
7.4: Areas of Tails of Distributions

Learning Objectives

- To learn how to find, for a normal random variable X and an area a, the value x^* of X so that $P(X < x^*) = a$ or that $P(X > x^*) = a$, whichever is required.

Definition: Left and Right Tails

The left tail of a density curve $y = f(x)$ of a continuous random variable X cut off by a value x^* of X is the region under the curve that is to the left of x^*, as shown by the shading in Figure 7.4.1(a). The right tail cut off by x^* is defined similarly, as indicated by the shading in Figure 7.4.1(b).



(a) (b)

Figure 7.4.1: Right and Left Tails of a Distribution

The probabilities tabulated in Figure 5.3.1 are areas of left tails in the standard normal distribution.

Tails of the Standard Normal Distribution

At times it is important to be able to solve the kind of problem illustrated by Figure 7.4.2. We have a certain specific area in mind, in this case the area 0.0125 of the shaded region in the figure, and we want to find the value z^* of Z that produces it. This is exactly the reverse of the kind of problems encountered so far. Instead of knowing a value z^* of Z and finding a corresponding area, we know the area and want to find z^*. In the case at hand, in the terminology of the definition just above, we wish to find the value z^* that cuts off a left tail of area 0.0125 in the standard normal distribution.

The idea for solving such a problem is fairly simple, although sometimes its implementation can be a bit complicated. In a nutshell, one reads the cumulative probability table for Z in reverse, looking up the relevant area in the interior of the table and reading off the value of Z from the margins.

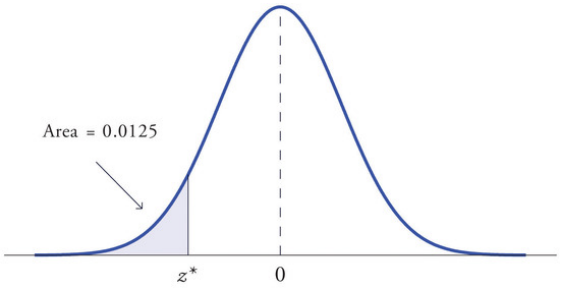


Figure 7.4.2: Z Value that Produces a Known Area

Example 7.4.1

Find the value z^* of Z as determined by Figure 7.4.2: the value z^* that cuts off a left tail of area 0.0125 in the standard normal distribution. In symbols, find the number z^* such that $P(Z < z^*) = 0.0125$.

Solution

The number that is known, 0.0125, is the area of a left tail, and as already mentioned the probabilities tabulated in Figure 5.3.1 are areas of left tails. Thus to solve this problem we need only search in the interior of Figure 5.3.1 for the number 0.0125. It lies in the row with the heading -2.2 and in the column with the heading 0.04. This means that $P(Z < -2.24) = 0.0125$, hence $z^* = -2.24$.

Example 7.4.2

Find the value z^* of Z as determined by Figure 7.4.3: the value z^* that cuts off a right tail of area 0.0250 in the standard normal distribution. In symbols, find the number z^* such that $P(Z > z^*) = 0.0250$.

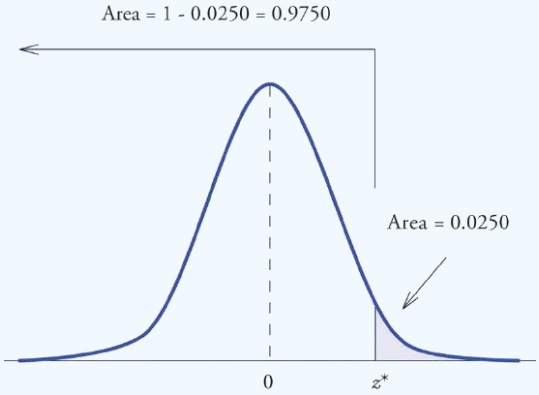


Figure 7.4.3: Z Value that Produces a Known Area

Solution

The important distinction between this example and the previous one is that here it is the area of a right tail that is known. In order to be able to use Figure 5.3.1 we must first find that area of the left tail cut off by the unknown number z^*. Since the total area under the density curve is 1, that area is $1 - 0.0250 = 0.9750$. This is the number we look for in the interior of Figure 5.3.1. It lies in the row with the heading 1.9 and in the column with the heading 0.06. Therefore $z^* = 1.96$.

Definition: standard normal random variable

The value of the standard normal random variable Z that cuts off a right tail of area c is denoted z_c. By symmetry, value of Z that cuts off a left tail of area c is $-z_c$. See Figure 7.4.4.

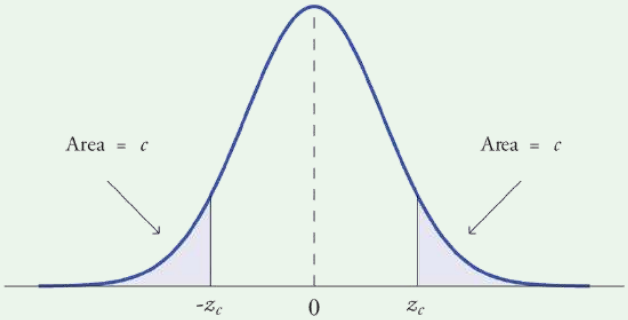


Figure 7.4.4: The Numbers z_c and $-z_c$

The previous two examples were atypical because the areas we were looking for in the interior of Figure 5.3.1 were actually there. The following example illustrates the situation that is more common.

Example 7.4.3

Find $z_{.01}$ and $-z_{.01}$, the values of Z that cut off right and left tails of area 0.01 in the standard normal distribution.

Solution

Since $-z_{.01}$ cuts off a left tail of area 0.01 and Figure 5.3.1 is a table of left tails, we look for the number 0.0100 in the interior of the table. It is not there, but falls between the two numbers 0.0102 and 0.0099 in the row with heading -2.3. The number 0.0099 is closer to 0.0100 than 0.0102 is, so for the hundredths place in $-z_{.01}$ we use the heading of the column that contains 0.0099, namely, 0.03, and write $-z_{.01} \approx -2.33$.

The answer to the second half of the problem is automatic: since $-z_{.01} = -2.33$, we conclude immediately that $z_{.01} = 2.33$.

We could just as well have solved this problem by looking for $z_{.01}$ first, and it is instructive to rework the problem this way. To begin with, we must first subtract 0.01 from 1 to find the area $1 - 0.0100 = 0.9900$ of the left tail cut off by the unknown number $z_{.01}$. See Figure 7.4.5. Then we search for the area 0.9900 in Figure 7.4.5. It is not there, but falls between the numbers 0.9898 and 0.9901 in the row with heading 2.3. Since 0.9901 is closer to 0.9900 than 0.9898 is, we use the column heading above it, 0.03, to obtain the approximation $z_{.01} \approx 2.33$. Then finally $-z_{.01} \approx -2.33$.

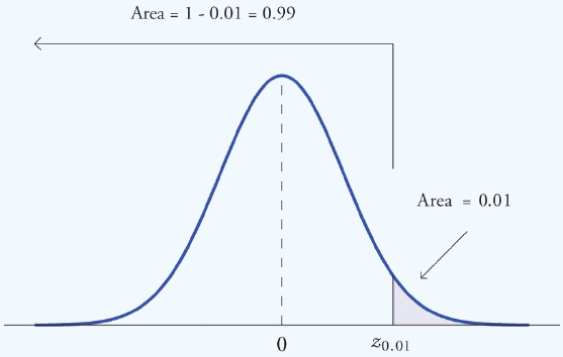


Figure 7.4.5: Computation of the Number $z_{.01}$

Tails of General Normal Distributions

The problem of finding the value x^* of a general normally distributed random variable X that cuts off a tail of a specified area also arises. This problem may be solved in two steps.

1. Suppose X is a normally distributed random variable with mean μ and standard deviation σ. To find the value x^* of X that cuts off a left or right tail of area c in the distribution of X:
2. find the value z^* of Z that cuts off a left or right tail of area c in the standard normal distribution; z^* is the z-score of x^*; compute x^* using the destandardization formula

$$x^* = \mu + z^*\sigma$$

In short, solve the corresponding problem for the standard normal distribution, thereby obtaining the z-score of x^*, then destandardize to obtain x^*.

Example 7.4.4

Find x^* such that $P(X < x^*) = 0.9332$, where X is a normal random variable with mean $\mu = 10$ and standard deviation $\sigma = 2.5$.

Solution

All the ideas for the solution are illustrated in Figure 7.4.6. Since 0.9332 is the area of a left tail, we can find z^* simply by looking for 0.9332 in the interior of Figure 5.3.1. It is in the row and column with headings 1.5 and 0.00, hence $z^* = 1.50$. Thus x^* is 1.50 standard deviations above the mean, so

$$x^* = \mu + z^*\sigma = 10 + (1.50)\cdot(0.2) = 13.75$$

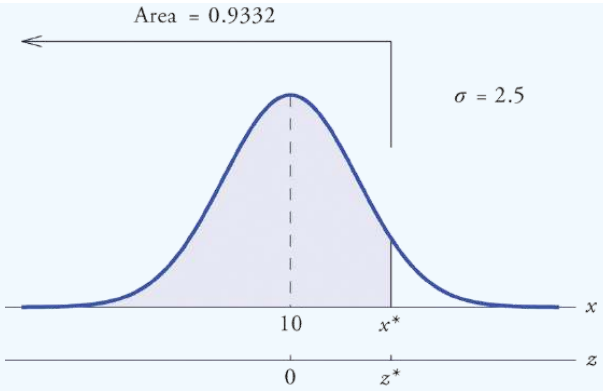


Figure 7.4.6: Tail of a Normally Distributed Random Variable

Example 7.4.5

Find x^* such that $P(X > x^*) = 0.65$, where X is a normal random variable with mean $\mu = 175$ and standard deviation $\sigma = 12$.

Solution

The situation is illustrated in Figure 7.4.7. Since 0.65 is the area of a right tail, we first subtract it from 1 to obtain $1 - 0.65 = 0.35$, the area of the complementary left tail. We find z^* by looking for 0.3500 in the interior of Figure 5.3.1. It is not present, but lies between table entries 0.3520 and 0.3483. The entry 0.3483 with row and column headings -0.3 and 0.09 is closer to 0.3500 than the other entry is, so $z^* \approx -0.39$. Thus x^* is 0.39 standard deviations below the mean, so

$$x^* = \mu + z^*\sigma = 175 + (-0.39) \cdot (12) = 170.32$$

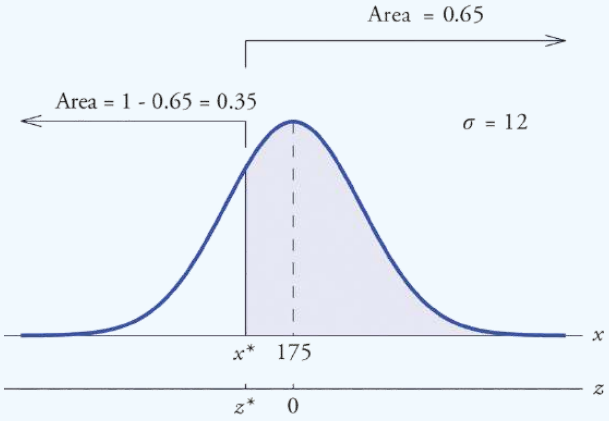


Figure 7.4.7: Tail of a Normally Distributed Random Variable

Example 7.4.6

Scores on a standardized college entrance examination (CEE) are normally distributed with mean 510 and standard deviation 60. A selective university decides to give serious consideration for admission to applicants whose CEE scores are in the top 5% of all CEE scores. Find the minimum score that meets this criterion for serious consideration for admission.

Solution

Let X denote the score made on the CEE by a randomly selected individual. Then X is normally distributed with mean 510 and standard deviation 60. The probability that X lie in a particular interval is the same as the proportion of all exam scores that lie in that interval. Thus the minimum score that is in the top 5% of all CEE is the score x^* that cuts off a right tail in the distribution of X of area 0.05 (5% expressed as a proportion). See Figure 7.4.8.

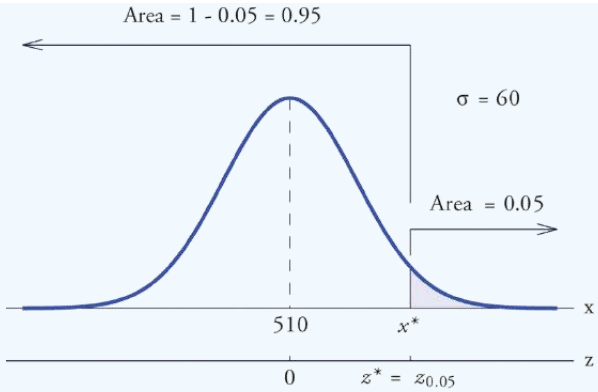


Figure 7.4.8: Tail of a Normally Distributed Random Variable

Since 0.0500 is the area of a right tail, we first subtract it from 1 to obtain $1-0.0500=0.9500$, the area of the complementary left tail. We find $z^*=z_{.05}$ by looking for 0.9500 in the interior of Figure 5.3.1. It is not present, and lies exactly half-way between the two nearest entries that are, 0.9495 and 0.9505. In the case of a tie like this, we will always average the values of Z corresponding to the two table entries, obtaining here the value $z^*=1.645$. Using this value, we conclude that x^* is 1.645 standard deviations above the mean, so

$$x^*=\mu+z^*\sigma=510+(1.645)\cdot(60)=608.7$$

Example 7.4.7

All boys at a military school must run a fixed course as fast as they can as part of a physical examination. Finishing times are normally distributed with mean 29 minutes and standard deviation 2 minutes. The middle 75% of all finishing times are classified as "average." Find the range of times that are average finishing times by this definition.

Solution

Let X denote the finish time of a randomly selected boy. Then X is normally distributed with mean 29 and standard deviation 2. The probability that X lie in a particular interval is the same as the proportion of all finish times that lie in that interval. Thus the situation is as shown in Figure $7.4.9$. Because the area in the middle corresponding to "average" times is 0.75, the areas of the two tails add up to $1-0.75=0.25$ in all. By the symmetry of the density curve each tail must have half of this total, or area 0.125 each. Thus the fastest time that is "average" has z-score $-z_{.125}$, which by Figure 5.3.1 is -1.15, and the slowest time that is "average" has z-score $z_{.125}=1.15$. The fastest and slowest times that are still considered average are

$$x_{fast}=\mu+(-z_{.125})\sigma=29+(-1.15)\cdot(2)=26.7$$

and

$$x_{slow}=\mu+z_{.125}\sigma=29+(1.15)\cdot(2)=31.3$$

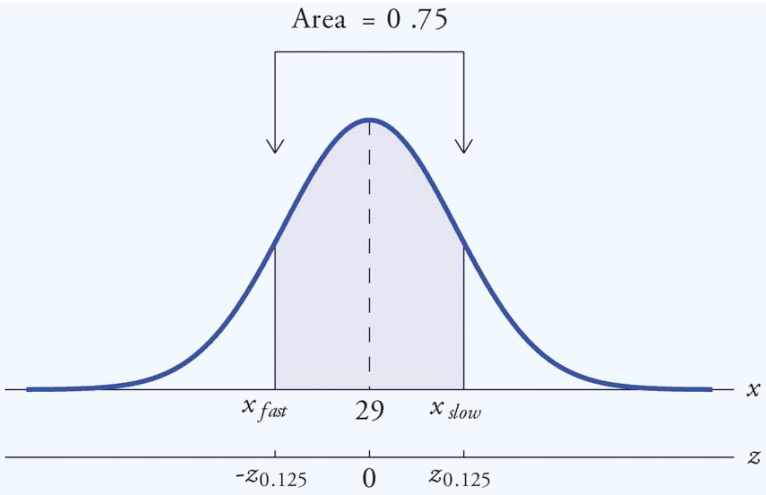


Figure 7.4.9: Distribution of Times to Run a Course

A boy has an average finishing time if he runs the course with a time between 26.7 and 31.3 minutes, or equivalently between 26 minutes 42 seconds and 31 minutes 18 seconds.

Key Takeaways

- The problem of finding the number z^* so that the probability $P(Z < z^*)$ is a specified value c is solved by looking for the number c in the interior of Figure 5.3.1 and reading z^* from the margins.
- The problem of finding the number z^* so that the probability $P(Z > z^*)$ is a specified value c is solved by looking for the complementary probability $1 - c$ in the interior of Figure 5.3.1 and reading z^* from the margins.
- For a normal random variable X with mean μ and standard deviation σ, the problem of finding the number x^* so that $P(X < x^*)$ is a specified value c (or so that $P(X > x^*)$ is a specified value c) is solved in two steps:
  - (1) solve the corresponding problem for Z with the same value of c, thereby obtaining the z-score, z^*, of x^*;
  - (2) find x^* using $x^* = \mu + z^*\sigma$.
- The value of Z that cuts off a right tail of area c in the standard normal distribution is denoted z_c.

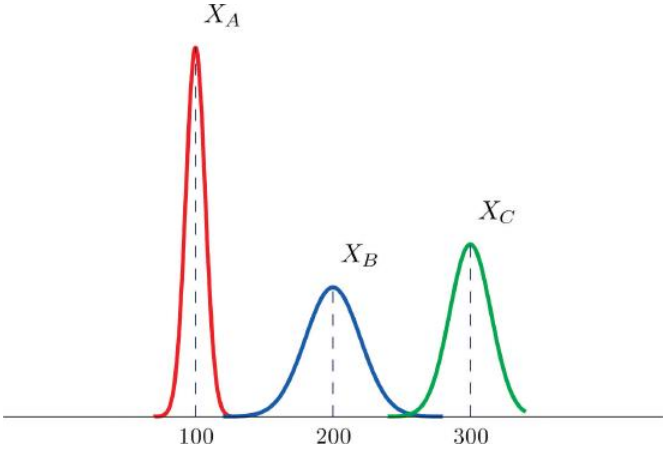
7.E: Continuous Random Variables (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

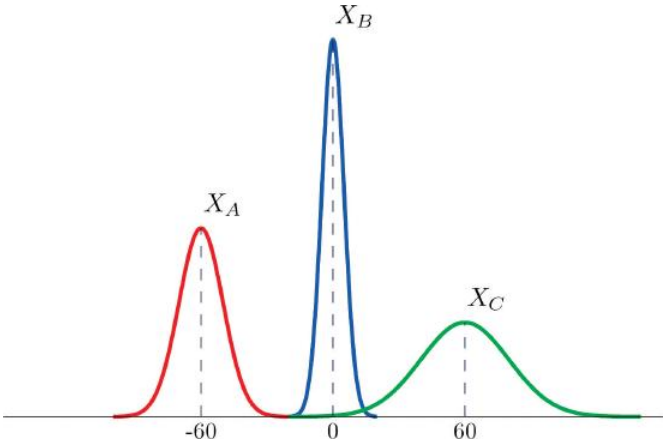
5.1: Continuous Random Variables

Basic

1. A continuous random variable X has a uniform distribution on the interval $[5, 12]$. Sketch the graph of its density function.
2. A continuous random variable X has a uniform distribution on the interval $[-3, 3]$. Sketch the graph of its density function.
3. A continuous random variable X has a normal distribution with mean 100 and standard deviation 10. Sketch a qualitatively accurate graph of its density function.
4. A continuous random variable X has a normal distribution with mean 73 and standard deviation 2.5. Sketch a qualitatively accurate graph of its density function.
5. A continuous random variable X has a normal distribution with mean 73. The probability that X takes a value greater than 80 is 0.212. Use this information and the symmetry of the density function to find the probability that X takes a value less than 66. Sketch the density curve with relevant regions shaded to illustrate the computation.
6. A continuous random variable X has a normal distribution with mean 169. The probability that X takes a value greater than 180 is 0.17. Use this information and the symmetry of the density function to find the probability that X takes a value less than 158. Sketch the density curve with relevant regions shaded to illustrate the computation.
7. A continuous random variable X has a normal distribution with mean 50.5. The probability that X takes a value less than 54 is 0.76. Use this information and the symmetry of the density function to find the probability that X takes a value greater than 47. Sketch the density curve with relevant regions shaded to illustrate the computation.
8. A continuous random variable X has a normal distribution with mean 12.25. The probability that X takes a value less than 13 is 0.82. Use this information and the symmetry of the density function to find the probability that X takes a value greater than 11.50. Sketch the density curve with relevant regions shaded to illustrate the computation.
9. The figure provided shows the density curves of three normally distributed random variables X_A, X_B and X_C. Their standard deviations (in no particular order) are 15, 7, and 20. Use the figure to identify the values of the means μ_A, μ_B, and μ_C and standard deviations σ_A, σ_B, and σ_C of the three random variables.



10. The figure provided shows the density curves of three normally distributed random variables X_A, X_B and X_C. Their standard deviations (in no particular order) are 20, 5, and 10. Use the figure to identify the values of the means μ_A, μ_B, and μ_C and standard deviations σ_A, σ_B, and σ_C of the three random variables.



Applications

11. Dogberry's alarm clock is battery operated. The battery could fail with equal probability at any time of the day or night. Every day Dogberry sets his alarm for $6:30\ a.m.$ and goes to bed at $10:00\ p.m.$. Find the probability that when the clock battery finally dies, it will do so at the most inconvenient time, between $10:00\ p.m.$ and $6:30\ a.m.$.
12. Buses running a bus line near Desdemona's house run every 15 minutes. Without paying attention to the schedule she walks to the nearest stop to take the bus to town. Find the probability that she waits more than 10 minutes.
13. The amount X of orange juice in a randomly selected half-gallon container varies according to a normal distribution with mean 64 ounces and standard deviation 0.25 ounce.
    a. Sketch the graph of the density function for X.
    b. What proportion of all containers contain less than a half gallon (64 ounces)? Explain.
    c. What is the median amount of orange juice in such containers? Explain.
14. The weight X of grass seed in bags marked 50 lb varies according to a normal distribution with mean 50 lb and standard deviation 1 ounce (0.0625lb).
    a. Sketch the graph of the density function for X.
    b. What proportion of all bags weigh less than 50 pounds? Explain.
    c. What is the median weight of such bags? Explain.

Answers

1. The graph is a horizontal line with height $1/7$ from $x = 5$ to $x = 12$
2.
3. The graph is a bell-shaped curve centered at 100 and extending from about 70 to 130.
4.
5. 0.212
6.
7. 0.76
8.
9. $\mu_A = 100,\ \mu_B = 200,\ \mu_C = 300,\ \sigma_A = 7,\ \sigma_B = 20,\ \sigma_C = 15$
10.
11. 0.3542
12.
13. a. The graph is a bell-shaped curve centered at 64 and extending from about 63.25 to 64.75.
    b. 0.5
    c. 64

5.2: The Standard Normal Distribution

Basic

1. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated./* <![CDATA[*/7.E.5/*]]>*/
   a. $P(Z < -1.72)$
   b. $P(Z < 2.05)$
   c. $P(Z < 0)$
   d. $P(Z > -2.11)$
   e. $P(Z > 1.63)$
   f. $P(Z > 2.36)$
2. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated./* <![CDATA[*/7.E.5/*]]>*/
   a. $P(Z < -1.17)$
   b. $P(Z < -0.05)$
   c. $P(Z < 0.66)$
   d. $P(Z > -2.43)$
   e. $P(Z > -1.00)$
   f. $P(Z > 2.19)$
3. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(-2.15 < Z < -1.09)$
   b. $P(-0.93 < Z < 0.55)$
   c. $P(0.68 < Z < 2.11)$
4. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(-1.99 < Z < -1.03)$
   b. $P(-0.87 < Z < 1.58)$
   c. $P(0.33 < Z < 0.96)$
5. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(-4.22 < Z < -1.39)$
   b. $P(-1.37 < Z < 5.11)$
   c. $P(Z < -4.31)$
   d. $P(Z < 5.02)$
6. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(Z > -5.31)$
   b. $P(-4.08 < Z < 0.58)$
   c. $P(Z < -6.16)$
   d. $P(-0.51 < Z < 5.63)$
7. Use Figure 7.1.5: Cumulative Normal Probability to find the probability listed. Find the second probability without referring to the table, but using the symmetry of the standard normal density curve instead. Sketch the density curve with relevant regions shaded to illustrate the computation.
   a. $P(Z < -1.08),\ P(Z > 1.08)$
   b. $P(Z < -0.36),\ P(Z > 0.36)$
   c. $P(Z < 1.25),\ P(Z > -1.25)$
   d. $P(Z < 2.03),\ P(Z > -2.03)$
8. Use Figure 7.1.5: Cumulative Normal Probability to find the probability listed. Find the second probability without referring to the table, but using the symmetry of the standard normal density curve instead. Sketch the density curve with relevant regions shaded to illustrate the computation.
   a. $P(Z < -2.11),\ P(Z > 2.11)$
   b. $P(Z < -0.88),\ P(Z > 0.88)$
   c. $P(Z < 2.44),\ P(Z > -2.44)$
   d. $P(Z < 3.07),\ P(Z > -3.07)$
9. The probability that a standard normal random variable Z takes a value in the union of intervals $(-\infty, -\alpha] \cup [\alpha, \infty)$, which arises in applications, will be denoted $P(Z \leq -a \text{ or } Z \geq a)$. Use Figure 7.1.5: Cumulative Normal Probability to find the

following probabilities of this type. Sketch the density curve with relevant regions shaded to illustrate the computation. Because of the symmetry of the standard normal density curve you need to use Figure 7.1.5: Cumulative Normal Probability only one time for each part.

a. $P(Z < -1.29 \text{ or } Z > 1.29)$
b. $P(Z < -2.33 \text{ or } Z > 2.33)$
c. $P(Z < -1.96 \text{ or } Z > 1.96)$
d. $P(Z < -3.09 \text{ or } Z > 3.09)$

10. The probability that a standard normal random variable Z takes a value in the union of intervals $(-\infty, -\alpha] \cup [\alpha, \infty)$, which arises in applications, will be denoted $P(Z \leq -a \text{ or } Z \geq a)$. Use Figure 7.1.5: Cumulative Normal Probability to find the following probabilities of this type. Sketch the density curve with relevant regions shaded to illustrate the computation. Because of the symmetry of the standard normal density curve you need to use Figure 7.1.5: Cumulative Normal Probability only one time for each part.

a. $P(Z < -2.58 \text{ or } Z > 2.58)$
b. $P(Z < -2.81 \text{ or } Z > 2.81)$
c. $P(Z < -1.65 \text{ or } Z > 1.65)$
d. $P(Z < -2.43 \text{ or } Z > 2.43)$

Answers

1. a. 0.0427
   b. 0.9798
   c. 0.5
   d. 0.9826
   e. 0.0516
   f. 0.0091
2.
3. a. 0.1221
   b. 0.5326
   c. 0.2309
4.
5. a. 0.0823
   b. 0.9147
   c. 0.0000
   d. 1.0000
6.
7. a. $0.1401,\ 0.1401$
   b. $0.3594,\ 0.3594$
   c. $0.8944,\ 0.8944$
   d. $0.9788,\ 0.9788$
8.
9. a. 0.1970
   b. 0.01980
   c. 0.0500
   d. 0.0020

5.3: Probability Computations for General Normal Random Variables

Basic

1. X is a normally distributed random variable with mean 57 and standard deviation 6. Find the probability indicated.

a. $P(X < 59.5)$
b. $P(X < 46.2)$
c. $P(X > 52.2)$

d. $P(X > 70)$

2. X is a normally distributed random variable with mean -25 and standard deviation 4. Find the probability indicated.
   a. $P(X < -27.2)$
   b. $P(X < -14.8)$
   c. $P(X > -33.1)$
   d. $P(X > -16.5)$
3. X is a normally distributed random variable with mean 112 and standard deviation 15. Find the probability indicated.
   a. $P(100 < X < 125)$
   b. $P(91 < X < 107)$
   c. $P(118 < X < 160)$
4. X is a normally distributed random variable with mean 72 and standard deviation 22. Find the probability indicated.
   a. $P(78 < X < 127)$
   b. $P(60 < X < 90)$
   c. $P(49 < X < 71)$
5. X is a normally distributed random variable with mean 500 and standard deviation 25. Find the probability indicated.
   a. $P(X < 400)$
   b. $P(466 < X < 625)$
6. X is a normally distributed random variable with mean 0 and standard deviation 0.75. Find the probability indicated.
   a. $P(-4.02 < X < 3.82)$
   b. $P(X > 4.11)$
7. X is a normally distributed random variable with mean 15 and standard deviation 1. Use Figure 7.1.5 $/* <![CDATA[*/7.E.5/*]]>*/$ Cumulative Normal Probability to find the first probability listed. Find the second probability using the symmetry of the density curve. Sketch the density curve with relevant regions shaded to illustrate the computation.
   a. $P(X < 12),\ P(X > 18)$
   b. $P(X < 14),\ P(X > 16)$
   c. $P(X < 11.25),\ P(X > 18.75)$
   d. $P(X < 12.67),\ P(X > 17.33)$
8. X is a normally distributed random variable with mean 100 and standard deviation 10. Use Figure 7.1.5 $/* <![CDATA[*/7.E.5/*]]>*/$ Cumulative Normal Probability to find the first probability listed. Find the second probability using the symmetry of the density curve. Sketch the density curve with relevant regions shaded to illustrate the computation.
   a. $P(X < 80),\ P(X > 120)$
   b. $P(X < 75),\ P(X > 125)$
   c. $P(X < 84.55),\ P(X > 115.45)$
   d. $P(X < 77.42),\ P(X > 122.58)$
9. X is a normally distributed random variable with mean 67 and standard deviation 13. The probability that X takes a value in the union of intervals $(-\infty, 67-a] \cup [67+a, \infty)$ will be denoted $P(X \leq 67-a \text{ or } X \geq 67+a)$. Use Figure 7.1.5 $/* <![CDATA[*/7.E.5/*]]>*/$ Cumulative Normal Probability to find the following probabilities of this type. Sketch the density curve with relevant regions shaded to illustrate the computation. Because of the symmetry of the density curve you need to use Figure 7.1.5$/* <![CDATA[*/7.E.5/*]]>*/$ Cumulative Normal Probability only one time for each part.
   a. $P(X < 57 \text{ or } X > 77)$
   b. $P(X < 47 \text{ or } X > 87)$
   c. $P(X < 49 \text{ or } X > 85)$
   d. $P(X < 37 \text{ or } X > 97)$
10. X is a normally distributed random variable with mean 288 and standard deviation 6. The probability that X takes a value in the union of intervals $(-\infty, 288-a] \cup [288+a, \infty)$ will be denoted $P(X \leq 288-a \text{ or } X \geq 288+a)$. Use Figure 7.1.5 $/* <![CDATA[*/7.E.5/*]]>*/$ Cumulative Normal Probability to find the following probabilities of this type. Sketch the

density curve with relevant regions shaded to illustrate the computation. Because of the symmetry of the density curve you need to use Figure 7.1.5/* <![CDATA[*/7.E. 5/*]]> */: Cumulative Normal Probability only one time for each part.

a. $P(X < 278 \text{ or } X > 298)$
b. $P(X < 268 \text{ or } X > 308)$
c. $P(X < 273 \text{ or } X > 303)$
d. $P(X < 280 \text{ or } X > 296)$

Applications

11. The amount X of beverage in a can labeled 12 ounces is normally distributed with mean 12.1 ounces and standard deviation 0.05 ounce. A can is selected at random.
    a. Find the probability that the can contains at least 12 ounces.
    b. Find the probability that the can contains between 11.9 and 12.1 ounces.
12. The length of gestation for swine is normally distributed with mean 114 days and standard deviation 0.75 day. Find the probability that a litter will be born within one day of the mean of 114.
13. The systolic blood pressure X of adults in a region is normally distributed with mean 112 mm Hg and standard deviation 15 mm Hg. A person is considered "prehypertensive" if his systolic blood pressure is between 120 and 130 mm Hg. Find the probability that the blood pressure of a randomly selected person is prehypertensive.
14. Heights X of adult women are normally distributed with mean 63.7 inches and standard deviation 2.71 inches. Romeo, who is 69.25 inches tall, wishes to date only women who are shorter than he but within 4 inches of his height. Find the probability that the next woman he meets will have such a height.
15. Heights X of adult men are normally distributed with mean 69.1 inches and standard deviation 2.92 inches. Juliet, who is 63.25 inches tall, wishes to date only men who are taller than she but within 6 inches of her height. Find the probability that the next man she meets will have such a height.
16. A regulation hockey puck must weigh between 5.5 and 6 ounces. The weights X of pucks made by a particular process are normally distributed with mean 5.75 ounces and standard deviation 0.11 ounce. Find the probability that a puck made by this process will meet the weight standard.
17. A regulation golf ball may not weigh more than 1.620 ounces. The weights X of golf balls made by a particular process are normally distributed with mean 1.361 ounces and standard deviation 0.09 ounce. Find the probability that a golf ball made by this process will meet the weight standard.
18. The length of time that the battery in Hippolyta's cell phone will hold enough charge to operate acceptably is normally distributed with mean 25.6 hours and standard deviation 0.32 hour. Hippolyta forgot to charge her phone yesterday, so that at the moment she first wishes to use it today it has been 26 hours 18 minutes since the phone was last fully charged. Find the probability that the phone will operate properly.
19. The amount of non-mortgage debt per household for households in a particular income bracket in one part of the country is normally distributed with mean $\$28,350$ and standard deviation $\$3,425$. Find the probability that a randomly selected such household has between $\$20,000$ and $\$30,000$ in non-mortgage debt.
20. Birth weights of full-term babies in a certain region are normally distributed with mean 7.125 lb and standard deviation 1.290 lb. Find the probability that a randomly selected newborn will weigh less than 5.5 lb, the historic definition of prematurity.
21. The distance from the seat back to the front of the knees of seated adult males is normally distributed with mean 23.8 inches and standard deviation 1.22 inches. The distance from the seat back to the back of the next seat forward in all seats on aircraft flown by a budget airline is 26 inches. Find the proportion of adult men flying with this airline whose knees will touch the back of the seat in front of them.
22. The distance from the seat to the top of the head of seated adult males is normally distributed with mean 36.5 inches and standard deviation 1.39 inches. The distance from the seat to the roof of a particular make and model car is 40.5 inches. Find the proportion of adult men who when sitting in this car will have at least one inch of headroom (distance from the top of the head to the roof).

Additional Exercises

23. The useful life of a particular make and type of automotive tire is normally distributed with mean $57,500$ miles and standard deviation 950 miles.
    a. Find the probability that such a tire will have a useful life of between $57,000$ and $58,000$ miles.

b. Hamlet buys four such tires. Assuming that their lifetimes are independent, find the probability that all four will last between $57,000$ and $58,000$ miles. (If so, the best tire will have no more than $1,000$ miles left on it when the first tire fails.) Hint: There is a binomial random variable here, whose value of p comes from part (a).

24. A machine produces large fasteners whose length must be within 0.5 inch of 22 inches. The lengths are normally distributed with mean 22.0 inches and standard deviation 0.17 inch.
    a. Find the probability that a randomly selected fastener produced by the machine will have an acceptable length.
    b. The machine produces 20 fasteners per hour. The length of each one is inspected. Assuming lengths of fasteners are independent, find the probability that all 20 will have acceptable length. Hint: There is a binomial random variable here, whose value of p comes from part (a).
25. The lengths of time taken by students on an algebra proficiency exam (if not forced to stop before completing it) are normally distributed with mean 28 minutes and standard deviation 1.5 minutes.
    a. Find the proportion of students who will finish the exam if a 30-minute time limit is set.
    b. Six students are taking the exam today. Find the probability that all six will finish the exam within the 30-minute limit, assuming that times taken by students are independent. Hint: There is a binomial random variable here, whose value of p comes from part (a).
26. Heights of adult men between 18 and 34 years of age are normally distributed with mean 69.1 inches and standard deviation 2.92 inches. One requirement for enlistment in the military is that men must stand between 60 and 80 inches tall.
    a. Find the probability that a randomly elected man meets the height requirement for military service.
    b. Twenty-three men independently contact a recruiter this week. Find the probability that all of them meet the height requirement. Hint: There is a binomial random variable here, whose value of p comes from part (a).
27. A regulation hockey puck must weigh between 5.5 and 6 ounces. In an alternative manufacturing process the mean weight of pucks produced is 5.75 ounce. The weights of pucks have a normal distribution whose standard deviation can be decreased by increasingly stringent (and expensive) controls on the manufacturing process. Find the maximum allowable standard deviation so that at most 0.005 of all pucks will fail to meet the weight standard. (Hint: The distribution is symmetric and is centered at the middle of the interval of acceptable weights.)
28. The amount of gasoline X delivered by a metered pump when it registers 5 gallons is a normally distributed random variable. The standard deviation σ of X measures the precision of the pump; the smaller σ is the smaller the variation from delivery to delivery. A typical standard for pumps is that when they show that 5 gallons of fuel has been delivered the actual amount must be between 4.97 and 5.03 gallons (which corresponds to being off by at most about half a cup). Supposing that the mean of X is 5, find the largest that σ can be so that $P(4.97 < X < 5.03)$ is 1.0000 to four decimal places when computed using Figure 7.1.5: Cumulative Normal Probability which means that the pump is sufficiently accurate. (Hint: The z-score of 5.03 will be the smallest value of Z so that Figure 7.1.5: Cumulative Normal Probability gives $P(Z < z) = 1.0000$).

Answers

1. a. 0.6628
   b. 0.0359
   c. 0.7881
   d. 0.0150
2.
3. a. 0.5959
   b. 0.2899
   c. 0.3439
4.
5. a. 0.0000
   b. 0.9131
6.
7. a. $0.0013,\ 0.0013$
   b. $0.1587,\ 0.1587$
   c. $0.0001,\ 0.0001$
   d. $0.0099,\ 0.0099$

8.

9. a. 0.4412
   b. 0.1236
   c. 0.1676
   d. 0.0208

10.

11. a. 0.9772
    b. 0.5000

12.

13. 0.1830

14.

15. 0.4971

16.

17. 0.9980

18.

19. 0.6771

20.

21. 0.0359

22.

23. a. 0.4038
    b. 0.0266

24.

25. a. 0.9082
    b. 0.5612

26.

27. 0.089

5.4: Areas of Tails of Distributions

Basic

1. Find the value of $z*$ that yields the probability shown.
   a. $P(Z < z*) = 0.0075$
   b. $P(Z < z*) = 0.9850$
   c. $P(Z > z*) = 0.8997$
   d. $P(Z > z*) = 0.0110$
2. Find the value of $z*$ that yields the probability shown.
   a. $P(Z < z*) = 0.3300$
   b. $P(Z < z*) = 0.9901$
   c. $P(Z > z*) = 0.0055$
   d. $P(Z > z*) = 0.7995$
3. Find the value of $z*$ that yields the probability shown.
   a. $P(Z < z*) = 0.1500$
   b. $P(Z < z*) = 0.7500$
   c. $P(Z > z*) = 0.3333$
   d. $P(Z > z*) = 0.8000$
4. Find the value of $z*$ that yields the probability shown.
   a. $P(Z < z*) = 0.2200$
   b. $P(Z < z*) = 0.6000$
   c. $P(Z > z*) = 0.0750$
   d. $P(Z > z*) = 0.8200$

5. Find the indicated value of Z. (It is easier to find $-z_c$and negate it.)
   a. $Z_{0.025}$
   b. $Z_{0.20}$
6. Find the indicated value of Z. (It is easier to find $-z_c$and negate it.)
   a. $Z_{0.002}$
   b. $Z_{0.02}$
7. Find the value of $x*$ that yields the probability shown, where X is a normally distributed random variable X with mean 83 and standard deviation 4.
   a. $P(X < x*) = 0.8700$
   b. $P(X > x*) = 0.0500$
8. Find the value of $x*$ that yields the probability shown, where X is a normally distributed random variable X with mean 54 and standard deviation 12.
   a. $P(X < x*) = 0.0900$
   b. $P(X > x*) = 0.6500$
9. X is a normally distributed random variable X with mean 15 and standard deviation 0.25. Find the values X_L and X_R of X that are symmetrically located with respect to the mean of X and satisfy $P(X_L < X < X_R) = 0.80$. (Hint. First solve the corresponding problem for Z).
10. X is a normally distributed random variable X with mean 28 and standard deviation 3.7. Find the values X_L and X_R of X that are symmetrically located with respect to the mean of X and satisfy $P(X_L < X < X_R) = 0.65$. (Hint. First solve the corresponding problem for Z).

Applications

11. Scores on a national exam are normally distributed with mean 382 and standard deviation 26.
    a. Find the score that is the 50^{th} percentile.
    b. Find the score that is the 90^{th} percentile.
12. Heights of women are normally distributed with mean 63.7 inches and standard deviation 2.47 inches.
    a. Find the height that is the 10^{th} percentile.
    b. Find the height that is the 80^{th} percentile.
13. The monthly amount of water used per household in a small community is normally distributed with mean $7,069$ gallons and standard deviation 58 gallons. Find the three quartiles for the amount of water used.
14. The quantity of gasoline purchased in a single sale at a chain of filling stations in a certain region is normally distributed with mean 11.6 gallons and standard deviation 2.78 gallons. Find the three quartiles for the quantity of gasoline purchased in a single sale.
15. Scores on the common final exam given in a large enrollment multiple section course were normally distributed with mean 69.35 and standard deviation 12.93. The department has the rule that in order to receive an A in the course his score must be in the top 10% of all exam scores. Find the minimum exam score that meets this requirement.
16. The average finishing time among all high school boys in a particular track event in a certain state is 5 minutes 17 seconds. Times are normally distributed with standard deviation 12 seconds.
    a. The qualifying time in this event for participation in the state meet is to be set so that only the fastest 5% of all runners qualify. Find the qualifying time. (Hint: Convert seconds to minutes.)
    b. In the western region of the state the times of all boys running in this event are normally distributed with standard deviation 12 seconds, but with mean 5 minutes 22 seconds. Find the proportion of boys from this region who qualify to run in this event in the state meet.
17. Tests of a new tire developed by a tire manufacturer led to an estimated mean tread life of $67,350$ miles and standard deviation of $1,120$ miles. The manufacturer will advertise the lifetime of the tire (for example, a “$50,000$ mile tire”) using the largest value for which it is expected that 98% of the tires will last at least that long. Assuming tire life is normally distributed, find that advertised value.
18. Tests of a new light led to an estimated mean life of $1,321$ hours and standard deviation of 106 hours. The manufacturer will advertise the lifetime of the bulb using the largest value for which it is expected that 90% of the bulbs will last at least that long. Assuming bulb life is normally distributed, find that advertised value.

19. The weights X of eggs produced at a particular farm are normally distributed with mean 1.72 ounces and standard deviation 0.12 ounce. Eggs whose weights lie in the middle 75% of the distribution of weights of all eggs are classified as “medium.” Find the maximum and minimum weights of such eggs. (These weights are endpoints of an interval that is symmetric about the mean and in which the weights of 75% of the eggs produced at this farm lie.)
20. The lengths X of hardwood flooring strips are normally distributed with mean 28.9 inches and standard deviation 6.12 inches. Strips whose lengths lie in the middle 80% of the distribution of lengths of all strips are classified as “average-length strips.” Find the maximum and minimum lengths of such strips. (These lengths are endpoints of an interval that is symmetric about the mean and in which the lengths of 80% of the hardwood strips lie.)
21. All students in a large enrollment multiple section course take common in-class exams and a common final, and submit common homework assignments. Course grades are assigned based on students' final overall scores, which are approximately normally distributed. The department assigns a C to students whose scores constitute the middle $2/3$ of all scores. If scores this semester had mean 72.5 and standard deviation 6.14, find the interval of scores that will be assigned a C.
22. Researchers wish to investigate the overall health of individuals with abnormally high or low levels of glucose in the blood stream. Suppose glucose levels are normally distributed with mean 96 and standard deviation 8.5 mg/dl, and that “normal” is defined as the middle 90% of the population. Find the interval of normal glucose levels, that is, the interval centered at 96 that contains 90% of all glucose levels in the population.

Additional Exercises

23. A machine for filling 2-liter bottles of soft drink delivers an amount to each bottle that varies from bottle to bottle according to a normal distribution with standard deviation 0.002 liter and mean whatever amount the machine is set to deliver.
    a. If the machine is set to deliver 2 liters (so the mean amount delivered is 2 liters) what proportion of the bottles will contain at least 2 liters of soft drink?
    b. Find the minimum setting of the mean amount delivered by the machine so that at least 99% of all bottles will contain at least 2 liters.
24. A nursery has observed that the mean number of days it must darken the environment of a species poinsettia plant daily in order to have it ready for market is 71 days. Suppose the lengths of such periods of darkening are normally distributed with standard deviation 2 days. Find the number of days in advance of the projected delivery dates of the plants to market that the nursery must begin the daily darkening process in order that at least 95% of the plants will be ready on time. (Poinsettias are so long-lived that once ready for market the plant remains salable indefinitely.)

Answers

1. a. -2.43
   b. 2.17
   c. -1.28
   d. 2.29
2.
3. a. -1.04
   b. 0.67
   c. 0.43
   d. -0.84
4.
5. a. 1.96
   b. 0.84
6.
7. a. 87.52
   b. 89.58
8.
9. 15.32
10.
11. a. 382
    b. 415

12.
13. $7030.14,\ 7069,\ 7107.86$
14.
15. 85.90
16.
17. $65,054$
18.
19. $1.58,\ 1.86$
20.
21. $66.5,\ 78.5$
22.
23. a. 0.5
    b. 2.005

Contributor

- Anonymous

CHAPTER OVERVIEW

8: Sampling Distributions

A statistic, such as the sample mean or the sample standard deviation, is a number computed from a sample. Since a sample is random, every statistic is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. As a random variable it has a mean, a standard deviation, and a probability distribution. The probability distribution of a statistic is called its sampling distribution. Typically sample statistics are not ends in themselves, but are computed in order to estimate the corresponding population parameters. This chapter introduces the concepts of the mean, the standard deviation, and the sampling distribution of a sample statistic, with an emphasis on the sample mean

8.1: The Mean and Standard Deviation of the Sample Mean

8.2: The Sampling Distribution of the Sample Mean

8.3: The Sample Proportion

8.4: Using the Central Limit Theorem

8.4E: Using the Central Limit Theorem (Exercises)

8.E: Sampling Distributions (Exercises)

8.1: The Mean and Standard Deviation of the Sample Mean

Learning Objectives

- To become familiar with the concept of the probability distribution of the sample mean.
- To understand the meaning of the formulas for the mean and standard deviation of the sample mean.

Suppose we wish to estimate the mean μ of a population. In actual practice we would typically take just one sample. Imagine however that we take sample after sample, all of the same size n, and compute the sample mean $\bar{x}$ each time. The sample mean x is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. We will write $\bar{X}$ when the sample mean is thought of as a random variable, and write x for the values that it takes. The random variable $\bar{X}$ has a mean, denoted $\mu_{\bar{X}}$, and a standard deviation, denoted $\sigma_{\bar{X}}$. Here is an example with such a small population and small sample size that we can actually write down every single sample.

Example 8.1.1

A rowing team consists of four rowers who weigh 152, 156, 160, and 164 pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\bar{X}$.

Solution

The following table shows all possible samples with replacement of size two, along with the mean of each:

Sample	Mean		Sample	Mean		Sample	Mean		Sample	Mean
152, 152	152		156, 152	154		160, 152	156		164, 152	158
152, 156	154		156, 156	156		160, 156	158		164, 156	160
152, 160	156		156, 160	158		160, 160	160		164, 160	162
152, 164	158		156, 164	160		160, 164	162		164, 164	164

The table shows that there are seven possible values of the sample mean $\bar{X}$. The value $\bar{x} = 152$ happens only one way (the rower weighing 152 pounds must be selected both times), as does the value $\bar{x} = 164$, but the other values happen more than one way, hence are more likely to be observed than 152 and 164 are. Since the 16 samples are equally likely, we obtain the probability distribution of the sample mean just by counting:

$\bar{x}$	152	154	156	158	160	162	164
$P(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Now we apply the formulas from Section 4.2 to $\bar{X}$. For $\mu_{\bar{X}}$, we obtain.

$$\begin{aligned}\mu_{\bar{X}} &= \sum \bar{x} P(\bar{x}) \\ &= 152\left(\frac{1}{16}\right) + 154\left(\frac{2}{16}\right) + 156\left(\frac{3}{16}\right) + 158\left(\frac{4}{16}\right) + 160\left(\frac{3}{16}\right) + 162\left(\frac{2}{16}\right) + 164\left(\frac{1}{16}\right) \\ &= 158\end{aligned}$$

For $\sigma_{\bar{X}}$, we first compute $\sum \bar{x}^2 P(\bar{x})$:

$$\sum \bar{x}^2 P(\bar{x}) = 152^2\left(\frac{1}{16}\right) + 154^2\left(\frac{2}{16}\right) + 156^2\left(\frac{3}{16}\right) + 158^2\left(\frac{4}{16}\right) + 160^2\left(\frac{3}{16}\right) + 162^2\left(\frac{2}{16}\right) + 164^2\left(\frac{1}{16}\right)$$

which is $24,974$, so that

$$\begin{aligned}\sigma_{\bar{x}} &= \sqrt{\sum \bar{x}^2 P(\bar{x}) - \mu_{\bar{x}}^2} \\ &= \sqrt{24,974 - 158^2} \\ &= \sqrt{10}\end{aligned}$$

The mean and standard deviation of the population $\{152, 156, 160, 164\}$ in the example are $\mu = 158$ and $\sigma = \sqrt{20}$. The mean of the sample mean $\bar{X}$ that we have just computed is exactly the mean of the population. The standard deviation of the sample mean $\bar{X}$ that we have just computed is the standard deviation of the population divided by the square root of the sample size: $\sqrt{10} = \sqrt{20}/\sqrt{2}$. These relationships are not coincidences, but are illustrations of the following formulas.

Definition: Sample mean and sample standard deviation

Suppose random samples of size n are drawn from a population with mean μ and standard deviation σ. The mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of the sample mean $\bar{X}$ satisfy

$$\mu_{\bar{X}} = \mu \tag{8.1.1}$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{8.1.2}$$

Equation 8.1.1 says that if we could take every possible sample from the population and compute the corresponding sample mean, then those numbers would center at the number we wish to estimate, the population mean μ. Equation 8.1.2 says that averages computed from samples vary less than individual measurements on the population do, and quantifies the relationship.

Example 8.1.2

The mean and standard deviation of the tax value of all vehicles registered in a certain state are $\mu = \$13,525$ and $\sigma = \$4,180$. Suppose random samples of size 100 are drawn from the population of vehicles. What are the mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of the sample mean $\bar{X}$?

Solution

Since $n = 100$, the formulas yield

$$\mu_{\bar{X}} = \mu = \$13,525$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4,180}{\sqrt{100}} = \$418$$

Key Takeaway

- The sample mean is a random variable; as such it is written $\bar{X}$, and $\bar{x}$ stands for individual values it takes.
- As a random variable the sample mean has a probability distribution, a mean $\mu_{\bar{X}}$, and a standard deviation $\sigma_{\bar{X}}$.
- There are formulas that relate the mean and standard deviation of the sample mean to the mean and standard deviation of the population from which the sample is drawn.

8.2: The Sampling Distribution of the Sample Mean

Learning Objectives

- To learn what the sampling distribution of $\overline{X}$ is when the sample size is large.
- To learn what the sampling distribution of $\overline{X}$ is when the population is normal.

In Example 6.1.1, we constructed the probability distribution of the sample mean for samples of size two drawn from the population of four rowers. The probability distribution is:

$\bar{x}$	152	154	156	158	160	162	164
$P(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Figure 8.2.1 shows a side-by-side comparison of a histogram for the original population and a histogram for this distribution. Whereas the distribution of the population is uniform, the sampling distribution of the mean has a shape approaching the shape of the familiar bell curve. This phenomenon of the sampling distribution of the mean taking on a bell shape even though the population distribution is not bell-shaped happens in general. Here is a somewhat more realistic example.

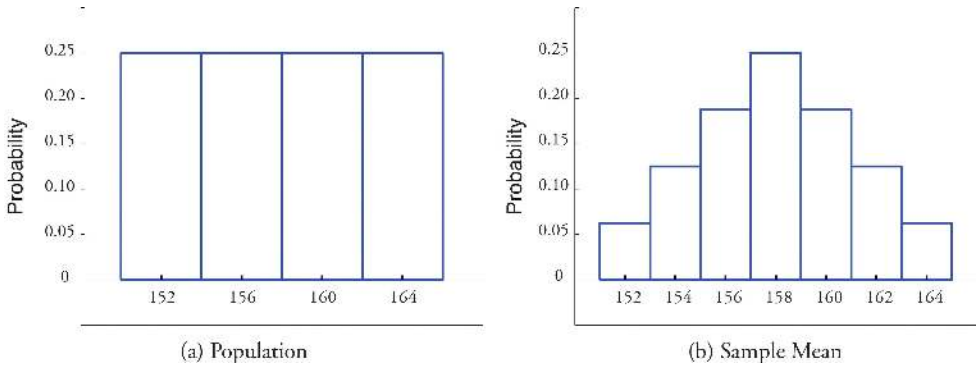


Figure 8.2.1: Distribution of a Population and a Sample Mean

Suppose we take samples of size 1, 5, 10, or 20 from a population that consists entirely of the numbers 0 and 1, half the population 0, half 1, so that the population mean is 0.5. The sampling distributions are:

$n = 1$:

$\bar{x}$	0	1
$P(\bar{x})$	0.5	0.5

$n = 5$:

$\bar{x}$	0	0.2	0.4	0.6	0.8	1
$P(\bar{x})$	0.03	0.16	0.31	0.31	0.16	0.03

$n = 10$:

$\bar{x}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$P(\bar{x})$	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00

$n = 20$:

$\bar{x}$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
$P(\bar{x})$	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.07	0.12	0.16	0.18

and

$\bar{x}$	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1
$P(\bar{x})$	0.16	0.12	0.07	0.04	0.01	0.00	0.00	0.00	0.00	0.00

Histograms illustrating these distributions are shown in Figure 8.2.2.

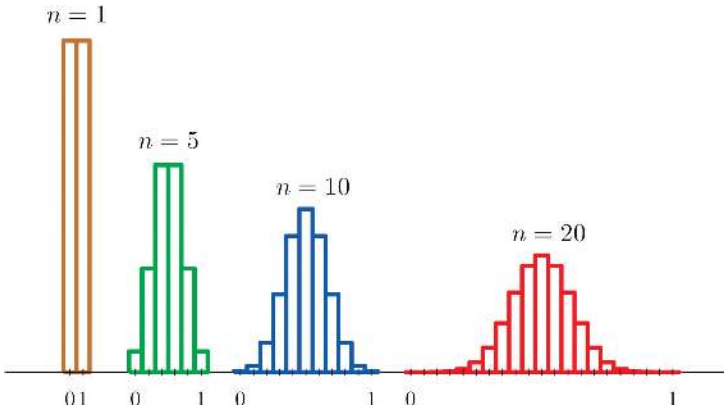


Figure 8.2.2: Distributions of the Sample Mean

As n increases the sampling distribution of $\overline{X}$ evolves in an interesting way: the probabilities on the lower and the upper ends shrink and the probabilities in the middle become larger in relation to them. If we were to continue to increase n then the shape of the sampling distribution would become smoother and more bell-shaped.

What we are seeing in these examples does not depend on the particular population distributions involved. In general, one may start with any distribution and the sampling distribution of the sample mean will increasingly resemble the bell-shaped normal curve as the sample size increases. This is the content of the Central Limit Theorem.

The Central Limit Theorem

For samples of size 30 or more, the sample mean is approximately normally distributed, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$, where n is the sample size. The larger the sample size, the better the approximation. The **Central Limit Theorem** is illustrated for several common population distributions in Figure 8.2.3.

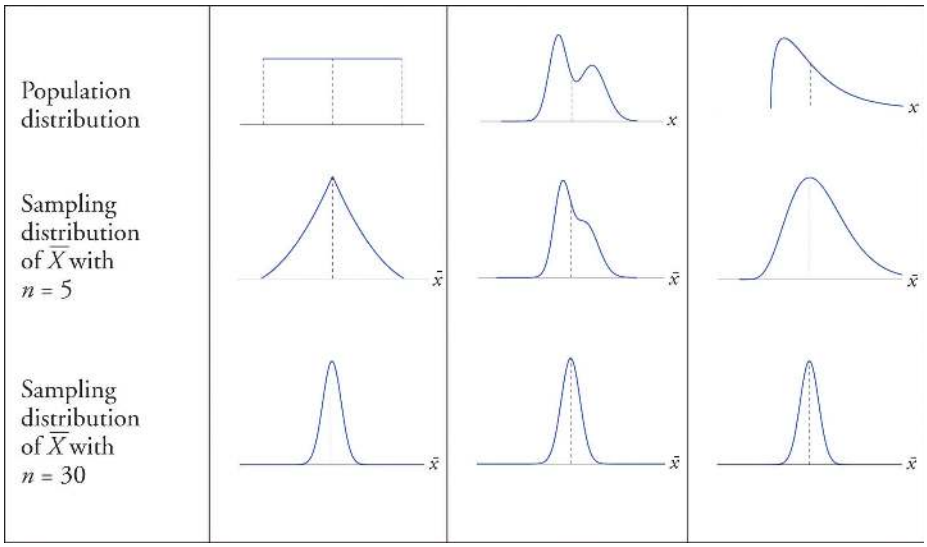


Figure 8.2.3: Distribution of Populations and Sample Means

The dashed vertical lines in the figures locate the population mean. Regardless of the distribution of the population, as the sample size is increased the shape of the sampling distribution of the sample mean becomes increasingly bell-shaped, centered on the population mean. Typically by the time the sample size is 30 the distribution of the sample mean is practically the same as a normal distribution.

The importance of the Central Limit Theorem is that it allows us to make probability statements about the sample mean, specifically in relation to its value in comparison to the population mean, as we will see in the examples. But to use the result properly we must first realize that there are two separate random variables (and therefore two probability distributions) at play:

1. X, the measurement of a single element selected at random from the population; the distribution of X is the distribution of the population, with mean the population mean μ and standard deviation the population standard deviation σ;
2. $\overline{X}$, the mean of the measurements in a sample of size n; the distribution of $\overline{X}$ is its sampling distribution, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$.

Example 8.2.1

Let $\overline{X}$ be the mean of a random sample of size 50 drawn from a population with mean 112 and standard deviation 40.

1. Find the mean and standard deviation of $\overline{X}$.
2. Find the probability that $\overline{X}$ assumes a value between 110 and 114.
3. Find the probability that $\overline{X}$ assumes a value greater than 113.

Solution

1. By the formulas in the previous section

$$\mu_{\overline{X}} = \mu = 112$$

and

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{50}} = 5.65685$$

2. Since the sample size is at least 30, the Central Limit Theorem applies: $\overline{X}$ is approximately normally distributed. We compute probabilities using Figure 5.3.1 in the usual way, just being careful to use $\sigma_{\overline{X}}$ and not σ when we standardize:

$$\begin{aligned} P(110 < \overline{X} < 114) &= P\left(\frac{110 - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < Z < \frac{114 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(\frac{110 - 112}{5.65685} < Z < \frac{114 - 112}{5.65685}\right) \\ &= P(-0.35 < Z < 0.35) \\ &= 0.6368 - 0.3632 \\ &= 0.2736 \end{aligned}$$

3. Similarly

$$\begin{aligned} P(\overline{X} > 113) &= P\left(Z > \frac{113 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z > \frac{113 - 112}{5.65685}\right) \\ &= P(Z > 0.18) \\ &= 1 - P(Z < 0.18) \\ &= 1 - 0.5714 \\ &= 0.4286 \end{aligned}$$

Note that if in the above example we had been asked to compute the probability that the value of a single randomly selected element of the population exceeds 113, that is, to compute the number $P(X > 113)$, we would not have been able to do so, since we do not know the distribution of X, but only that its mean is 112 and its standard deviation is 40. By contrast we could compute

$P(\overline{X} > 113)$ even without complete knowledge of the distribution of X because the Central Limit Theorem guarantees that $\overline{X}$ is approximately normal.

Example 8.2.2

The numerical population of grade point averages at a college has mean 2.61 and standard deviation 0.5. If a random sample of size 100 is taken from the population, what is the probability that the sample mean will be between 2.51 and 2.71?

Solution

The sample mean $\overline{X}$ has mean $\mu_{\overline{X}} = \mu = 2.61$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0.5}{10} = 0.05$, so

$$\begin{aligned} P(2.51 < \overline{X} < 2.71) &= P\left(\frac{2.51 - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < Z < \frac{2.71 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(\frac{2.51 - 2.61}{0.05} < Z < \frac{2.71 - 2.61}{0.05}\right) \\ &= P(-2 < Z < 2) \\ &= P(Z < 2) - P(Z < -2) \\ &= 0.9772 - 0.0228 \\ &= 0.9544 \end{aligned}$$

Normally Distributed Populations

The Central Limit Theorem says that no matter what the distribution of the population is, as long as the sample is "large," meaning of size 30 or more, the sample mean is approximately normally distributed. If the population is normal to begin with then the sample mean also has a normal distribution, regardless of the sample size.

> For samples of any size drawn from a normally distributed population, the sample mean is normally distributed, with mean $\mu_X = \mu$ and standard deviation $\sigma_X = \sigma/\sqrt{n}$, where n is the sample size.

The effect of increasing the sample size is shown in Figure 8.2.4.

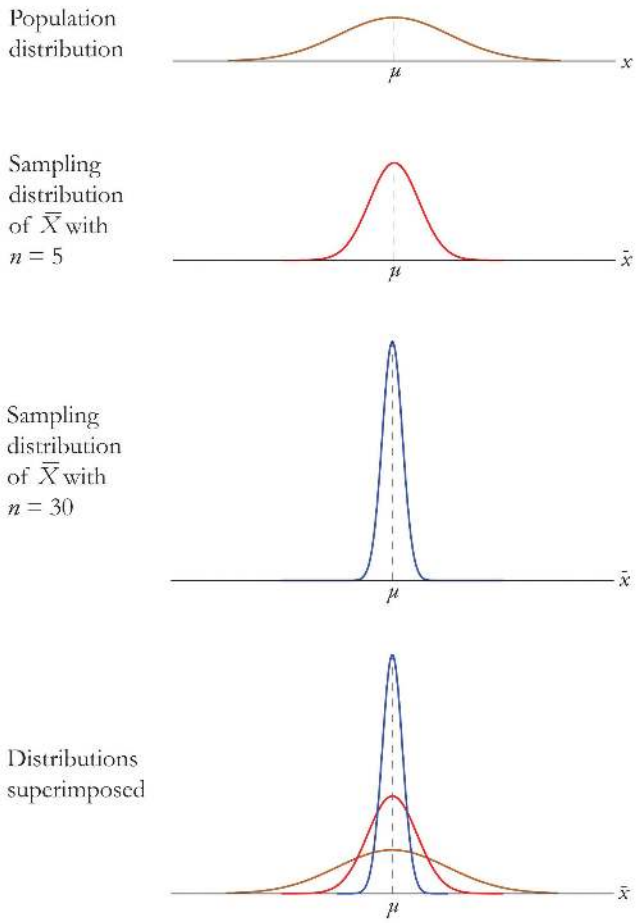


Figure 8.2.4: Distribution of Sample Means for a Normal Population

Example 8.2.3

A prototype automotive tire has a design life of $38,500$ miles with a standard deviation of $2,500$ miles. Five such tires are manufactured and tested. On the assumption that the actual population mean is $38,500$ miles and the actual population standard deviation is $2,500$ miles, find the probability that the sample mean will be less than $36,000$ miles. Assume that the distribution of lifetimes of such tires is normal.

Solution

For simplicity we use units of thousands of miles. Then the sample mean $\overline{X}$ has mean $\mu_{\overline{X}} = \mu = 38.5$ and standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{2.5}{\sqrt{5}} = 1.11803$. Since the population is normally distributed, so is $\overline{X}$, hence

$$\begin{aligned} P(\overline{X} < 36) &= P\left(Z < \frac{36 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z < \frac{36 - 38.5}{1.11803}\right) \\ &= P(Z < -2.24) \\ &= 0.0125 \end{aligned}$$

That is, if the tires perform as designed, there is only about a 1.25% chance that the average of a sample of this size would be so low.

Example 8.2.4

An automobile battery manufacturer claims that its midgrade battery has a mean life of 50 months with a standard deviation of 6 months. Suppose the distribution of battery lives of this particular brand is approximately normal.

a. On the assumption that the manufacturer's claims are true, find the probability that a randomly selected battery of this type will last less than 48 months.
b. On the same assumption, find the probability that the mean of a random sample of 36 such batteries will be less than 48 months.

Solution

a. Since the population is known to have a normal distribution

$$\begin{aligned} P(X<48) &= P\left(Z<\frac{48-\mu}{\sigma}\right) \\ &= P\left(Z<\frac{48-50}{6}\right) \\ &= P(Z<-0.33) \\ &= 0.3707 \end{aligned}$$

b. The sample mean has mean $\mu_{\overline{X}}=\mu=50$ and standard deviation $\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}=\frac{6}{\sqrt{36}}=1$. Thus

$$\begin{aligned} P(\overline{X}<48) &= P\left(Z<\frac{48-\mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z<\frac{48-50}{1}\right) \\ &= P(Z<-2) \\ &= 0.0228 \end{aligned}$$

Key Takeaway

- When the sample size is at least 30 the sample mean is normally distributed.
- When the population is normal the sample mean is normally distributed regardless of the sample size.

8.3: The Sample Proportion

Learning Objectives

- To recognize that the sample proportion $\hat{p}$ is a random variable.
- To understand the meaning of the formulas for the mean and standard deviation of the sample proportion.
- To learn what the sampling distribution of $\hat{p}$ is when the sample size is large.

Often sampling is done in order to estimate the proportion of a population that has a specific characteristic, such as the proportion of all items coming off an assembly line that are defective or the proportion of all people entering a retail store who make a purchase before leaving. The population proportion is denoted p and the sample proportion is denoted $\hat{p}$. Thus if in reality 43% of people entering a store make a purchase before leaving,

$$p = 0.43$$

if in a sample of 200 people entering the store, 78 make a purchase,

$$\hat{p} = \frac{78}{200} = 0.39.$$

The sample proportion is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. Viewed as a random variable it will be written $\hat{P}$. It has a mean $\mu_{\hat{P}}$ and a standard deviation $\sigma_{\hat{P}}$. Here are formulas for their values.

mean and standard deviation of the sample proportion

Suppose random samples of size n are drawn from a population in which the proportion with a characteristic of interest is p. The mean $\mu_{\hat{P}}$ and standard deviation $\sigma_{\hat{P}}$ of the sample proportion $\hat{P}$ satisfy

$$\mu_{\hat{P}} = p$$

and

$$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$$

where $q = 1 - p$.

The Central Limit Theorem has an analogue for the population proportion $\hat{p}$. To see how, imagine that every element of the population that has the characteristic of interest is labeled with a 1, and that every element that does not is labeled with a 0. This gives a numerical population consisting entirely of zeros and ones. Clearly the proportion of the population with the special characteristic is the proportion of the numerical population that are ones; in symbols,

$$p = \frac{\text{number of 1s}}{N}$$

But of course the sum of all the zeros and ones is simply the number of ones, so the mean μ of the numerical population is

$$\mu = \frac{\sum x}{N} = \frac{\text{number of 1s}}{N}$$

Thus the population proportion p is the same as the mean μ of the corresponding population of zeros and ones. In the same way the sample proportion $\hat{p}$ is the same as the sample mean $\bar{x}$. Thus the Central Limit Theorem applies to $\hat{p}$. However, the condition that the sample be large is a little more complicated than just being of size at least 30.

The Sampling Distribution of the Sample Proportion

For large samples, the sample proportion is approximately normally distributed, with mean $\mu_{\hat{P}} = p$ and standard deviation $\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$.

A sample is large if the interval $\left[p-3\sigma_{\hat{p}},\ p+3\sigma_{\hat{p}}\right]$ lies wholly within the interval $[0,1]$.

In actual practice p is not known, hence neither is $\sigma_{\hat{P}}$. In that case in order to check that the sample is sufficiently large we substitute the known quantity $\hat{p}$ for p. This means checking that the interval

$$\left[\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

lies wholly within the interval $[0,1]$. This is illustrated in the examples.

Figure 8.3.1 shows that when $p=0.1$, a sample of size 15 is too small but a sample of size 100 is acceptable.

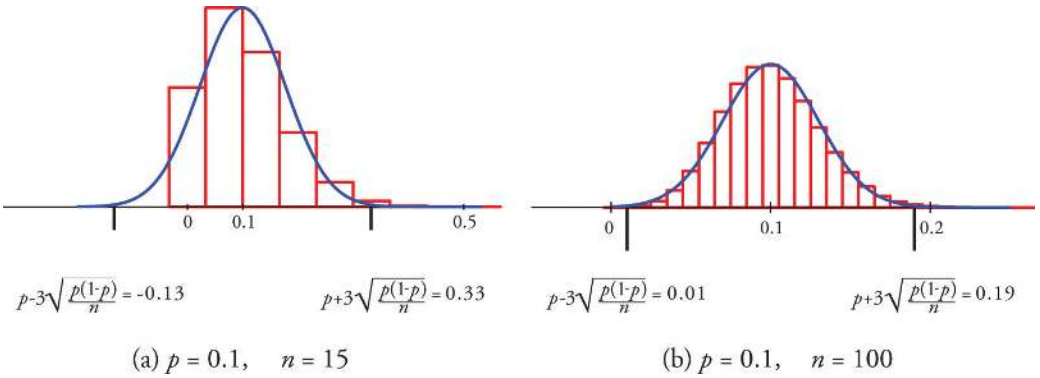


Figure 8.3.1: *Distribution of Sample Proportions*

Figure 8.3.2 shows that when $p=0.5$ a sample of size 15 is acceptable.

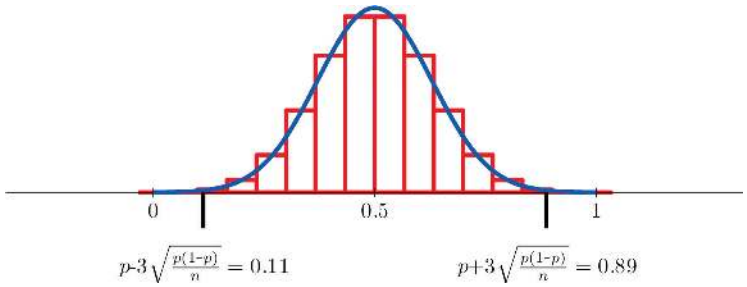


Figure 8.3.2: Distribution of Sample Proportions for $p=0.5$ *and* $n=15$

Example 8.3.1

Suppose that in a population of voters in a certain region 38% are in favor of particular bond issue. Nine hundred randomly selected voters are asked if they favor the bond issue.

1. Verify that the sample proportion $\hat{p}$ computed from samples of size 900 meets the condition that its sampling distribution be approximately normal.
2. Find the probability that the sample proportion computed from a sample of size 900 will be within 5 percentage points of the true population proportion.

Solution

1. The information given is that $p=0.38$, hence $q=1-p=0.62$. First we use the formulas to compute the mean and standard deviation of $\hat{p}$:

$$\mu_{\hat{p}}=p=0.38 \text{ and } \sigma_{\hat{P}}=\sqrt{\frac{pq}{n}}=\sqrt{\frac{(0.38)(0.62)}{900}}=0.01618$$

Then $3\sigma_{\hat{P}}=3(0.01618)=0.04854\approx 0.05$so

$$\left[\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]=[0.38-0.05,0.38+0.05]=[0.33,0.43]$$

which lies wholly within the interval $[0, 1]$, so it is safe to assume that $\hat{p}$ is approximately normally distributed.

2. To be within 5 percentage points of the true population proportion 0.38 means to be between $0.38 - 0.05 = 0.33$ and $0.38 + 0.05 = 0.43$. Thus

$$\begin{aligned} P(0.33 < \hat{P} < 0.43) &= P\left(\frac{0.33 - \mu_{\hat{P}}}{\sigma_{\hat{P}}} < Z < \frac{0.43 - \mu_{\hat{P}}}{\sigma_{\hat{P}}}\right) \\ &= P\left(\frac{0.33 - 0.38}{0.01618} < Z < \frac{0.43 - 0.38}{0.01618}\right) \\ &= P(-3.09 < Z < 3.09) \\ &= P(3.09) - P(-3.09) \\ &= 0.9990 - 0.0010 \\ &= 0.9980 \end{aligned}$$

Example 8.3.2

An online retailer claims that 90% of all orders are shipped within 12 hours of being received. A consumer group placed 121 orders of different sizes and at different times of day; 102 orders were shipped within 12 hours.

1. Compute the sample proportion of items shipped within 12 hours.
2. Confirm that the sample is large enough to assume that the sample proportion is normally distributed. Use $p = 0.90$, corresponding to the assumption that the retailer's claim is valid.
3. Assuming the retailer's claim is true, find the probability that a sample of size 121 would produce a sample proportion so low as was observed in this sample.
4. Based on the answer to part (c), draw a conclusion about the retailer's claim.

Solution

1. The sample proportion is the number x of orders that are shipped within 12 hours divided by the number n of orders in the sample:

$$\hat{p} = \frac{x}{n} = \frac{102}{121} = 0.84$$

2. Since $p = 0.90$, $q = 1 - p = 0.10$, and $n = 121$,

$$\sigma_{\hat{P}} = \sqrt{\frac{(0.90)(0.10)}{121}} = 0.0\overline{27}$$

hence

$$\left[p - 3\sigma_{\hat{P}},\ p + 3\sigma_{\hat{P}}\right] = [0.90 - 0.08, 0.90 + 0.08] = [0.82, 0.98]$$

Because

$$[0.82, 0.98] \subset [0, 1]$$

it is appropriate to use the normal distribution to compute probabilities related to the sample proportion $\hat{P}$.

3. Using the value of $\hat{P}$ from part (a) and the computation in part (b),

$$\begin{aligned} P(\hat{P} \leq 0.84) &= P\left(Z \leq \frac{0.84 - \mu_{\hat{P}}}{\sigma_{\hat{P}}}\right) \\ &= P\left(Z \leq \frac{0.84 - 0.90}{0.0\overline{27}}\right) \\ &= P(Z \leq -2.20) \\ &= 0.0139 \end{aligned}$$

4. The computation shows that a random sample of size 121 has only about a 1.4% chance of producing a sample proportion as the one that was observed, $\hat{p} = 0.84$, when taken from a population in which the actual proportion is 0.90. This is so unlikely that it is reasonable to conclude that the actual value of p is less than the 90% claimed.

Key Takeaway

- The sample proportion is a random variable $\hat{P}$.
- There are formulas for the mean $\mu_{\hat{P}}$, and standard deviation $\sigma_{\hat{P}}$ of the sample proportion.
- When the sample size is large the sample proportion is normally distributed.

8.4: Using the Central Limit Theorem

It is important for you to understand when to use the central limit theorem (clt). If you are being asked to find the probability of the mean, use the clt for the mean. If you are being asked to find the probability of a sum or total, use the clt for sums. This also applies to percentiles for means and sums.

> If you are being asked to find the probability of an individual value, do not use the clt. Use the distribution of its random variable.

Law of Large Numbers

The law of large numbers says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample tends to get closer and closer to μ. From the central limit theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\bar{x}$ must be close to the population mean μ. We can say that μ is the value that the sample means approach as n gets larger. The central limit theorem illustrates the law of large numbers.

Example 8.4.1

A study involving stress is conducted among the students on a college campus. The stress scores follow a uniform distribution with the lowest stress score equal to one and the highest equal to five. Using a sample of 75 students, find:

a. The probability that the **mean stress score** for the 75 students is less than two.
b. The 90th percentile for the **mean stress score** for the 75 students.
c. The probability that the **total of the 75 stress scores** is less than 200.
d. The 90th percentile for the **total stress score** for the 75 students.

Solutions

Let $X =$ one stress score.

Problems a and b ask you to find a probability or a percentile for a mean. Problems c and d ask you to find a probability or a percentile for a **total or sum**. The sample size, n, is equal to 75.

Since the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$ where $a = 1$ and $b = 5$.

$$\mu_x = \frac{a+b}{2} = \frac{1+5}{2} = 3 \tag{8.4.1}$$

$$\sigma_x = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15 \tag{8.4.2}$$

For problems 1. and 2., let $\bar{X} =$ the mean stress score for the 75 students. Then,

$$\bar{X} \sim N\left(3, \frac{1,15}{\sqrt{75}}\right) \tag{8.4.3}$$

where $n = 75$.

a. Find $P(\bar{x} < 2)$. Draw the graph.
b. Find the 90th percentile for the mean of 75 stress scores. Draw a graph.
c. Find $P(\sum x < 2000)$. Draw the graph.
d. Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Answers

a. $P(\bar{x} < 2) = 0$

The probability that the mean stress score is less than two is about zero.

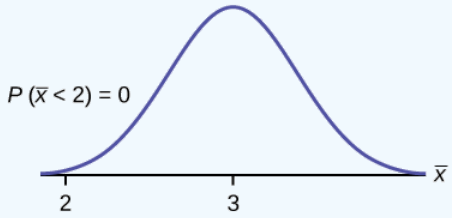


Figure 8.4.1.

normalcdf $\left(1, 2, 3, \frac{1.15}{\sqrt{75}}\right) = 0$

REMINDER

The smallest stress score is one

b. Let $k =$ the 90th percentile.

Find k, where $P(\bar{x} < k) = 0.90$.

$k = 3.2$

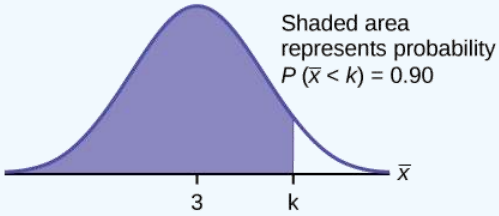


Figure 8.4.2.

The 90th percentile for the mean of 75 scores is about 3.2. This tells us that 90% of all the means of 75 stress scores are at most 3.2, and that 10% are at least 3.2.

invNorm $\left(0.90, 3, 1.\frac{1.15}{\sqrt{75}}\right) = 3.2$

For problems c and d, let $\sum X =$ the sum of the 75 stress scores. Then,

$$\sum X \sim N((75)(3), (\sqrt{75})(1.15)) \tag{8.4.4}$$

c. The mean of the sum of 75 stress scores is $(75)(3) = 225$

The standard deviation of the sum of 75 stress scores is $(\sqrt{75})(1.15) = 9.96$

$P(\sum x < 200)$

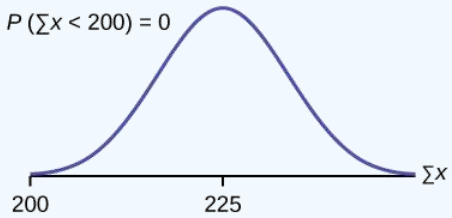


Figure 8.4.3.

The probability that the total of 75 scores is less than 200 is about zero.

normalcdf $75, 200, (75)(3), (\sqrt{75})(1.15)$

REMINDER

The smallest total of 75 stress scores is 75, because the smallest single score is one.

d. Let $k =$ the 90th percentile.

Find k where $P(\sum x < k) = 0.90$.

$k = 237.8$

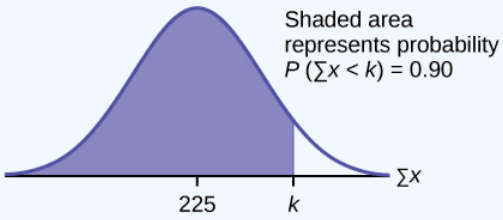


Figure 8.4.4.

The 90th percentile for the sum of 75 scores is about 237.8. This tells us that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

`invNorm` $\left(0.90, (75)(3), (\sqrt{75})(1.15)\right) = 237.8$

Exercise 8.4.1

Use the information in Example 8.4.1, but use a sample size of 55 to answer the following questions.

a. Find $P(\bar{x} < 7)$.
b. Find $P(\sum x < 7)$.
c. Find the 80th percentile for the mean of 55 scores.
d. Find the 85th percentile for the sum of 55 scores.

Answer

a. 0.0265
b. 0.2789
c. 3.13
d. 173.84

Example 8.4.2

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the *excess time used* follows an exponential distribution with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let $X =$ the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.

$X \sim Exp\left(\frac{1}{22}\right)$. From previous chapters, we know that $\mu = 22$ and $\sigma = 22$.

Let $\bar{X}$ = the mean excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.

$$\bar{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right) \tag{8.4.5}$$

by the central limit theorem for sample means

a. Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P(\bar{x} > 20)$. Draw the graph.
b. Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P(x > 20)$.
c. Explain why the probabilities in parts a and b are different.
d. Find the 95th percentile for the **sample mean excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Answer

a. Find: $P(\bar{x} > 20)$

$P(\bar{x} > 20) = 0.79199$ using `normalcdf` $\left(20, 1E99, 22, \frac{22}{\sqrt{80}}\right)$

The probability is 0.7919 that the mean excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.

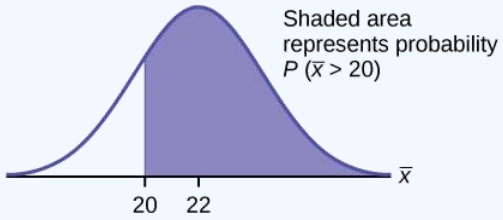


***Figure* 8.4.5.**

REMINDER

1E99 = 10^{99} and –1E99 = -10^{99}. Press the `EE` key for E. Or just use 10^{99} instead of 1E99.

b. Find $P(x > 20)$. Remember to use the exponential distribution for an **individual:** $X \sim Exp\left(\frac{1}{22}\right)$.

$P(x > 20) = e^{\left(-\left(\frac{1}{22}\right)(20)\right)}$ or $e^{(-0.04545(20))} = 0.4029$

c. i. $P(x > 20) = 0.4029$ but $P(\bar{x} > 20) = 0.7919$

ii. The probabilities are not equal because we use different distributions to calculate the probability for individuals and for means.

iii. When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the clt. Use the clt with the normal distribution when you are being asked to find the probability for a mean.

d. Let k = the 95th percentile. Find k where $P(\bar{x} < k) = 0.95$

$k = 26.0$ using `invNorm` $\left(0.95, 22, \frac{22}{\sqrt{80}}\right) = 26.0$

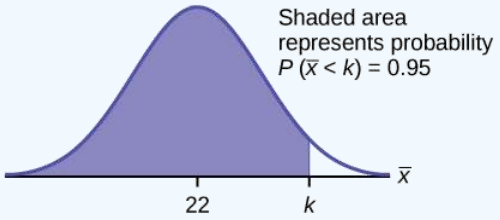


***Figure* 8.4.6.**

The 95th percentile for the **sample mean excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time.

Ninety five percent of such samples would have means under 26 minutes; only five percent of such samples would have means above 26 minutes.

? Exercise 8.4.2

Use the information in Example 8.4.2, but change the sample size to 144.

a. Find $P(20 < \bar{x} < 30)$.
b. Find $P(\sum x \text{ is at least } 3,000)$.
c. Find the 75th percentile for the sample mean excess time of 144 customers.

d. Find the 85th percentile for the sum of 144 excess times used by customers.

Answer

a. 0.8623
b. 0.7377
c. 23.2
d. 3,441.6

Example 8.4.3

In the United States, someone is sexually assaulted every two minutes, on average, according to a number of studies. Suppose the standard deviation is 0.5 minutes and the sample size is 100.

a. Find the median, the first quartile, and the third quartile for the sample mean time of sexual assaults in the United States.
b. Find the median, the first quartile, and the third quartile for the sum of sample times of sexual assaults in the United States.
c. Find the probability that a sexual assault occurs on the average between 1.75 and 1.85 minutes.
d. Find the value that is two standard deviations above the sample mean.
e. Find the *IQR* for the sum of the sample times.

Answer

a. We have, $\mu_x = \mu = 2$ and $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{10} = 0.05$. Therefore:
   a. 50th percentile $= \mu_x = \mu = 2$
   b. 25th percentile $= \text{invNorm}(0.25, 2, 0.05) = 1.97$
   c. 75th percentile $= \text{invNorm}(0.75, 2, 0.05) = 2.03$
b. We have $\mu_{\sum X} = n(\mu_x) = 100(2)$ and $\sigma_{\mu X} = \sqrt{n}(\sigma_x) = 10(0.5) = 5$. Therefore
   a. 50th percentile = $\mu_{\sum X} = n(\mu_X) = 100(2) = 200$
   b. 25th percentile $= \text{invNorm}(0.25, 200, 5) = 196.63$
   c. 75th percentile $= \text{invNorm}(0.75, 200, 5) = 203.37$
c. $P(1.75 < barx < 1.85) =$ `normalcdf` $(1.75, 1.85, 2, 0.05) = 0.0013$
d. Using the z-score equation, $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$, and solving for x, we have $x = 2(0.05) + 2 = 2.1$
e. The IQR is 75th percentile – 25th percentile $= 203.37 - 196.63 = 6.74$

Exercise 8.4.3

Based on data from the National Health Survey, women between the ages of 18 and 24 have an average systolic blood pressures (in mm Hg) of 114.8 with a standard deviation of 13.1. Systolic blood pressure for women between the ages of 18 to 24 follow a normal distribution.

a. If one woman from this population is randomly selected, find the probability that her systolic blood pressure is greater than 120.
b. If 40 women from this population are randomly selected, find the probability that their mean systolic blood pressure is greater than 120.
c. If the sample were four women between the ages of 18 to 24 and we did not know the original distribution, could the central limit theorem be used?

Answer

a. $P(x > 120) =$ `normalcdf` $(120, 99, 114.8, 13.1) = 0.0272$ There is about a 3%, that the randomly selected woman will have systolics blood pressure greater than 120.
b. $P(\bar{x} > 120) =$ `normalcdf` $\left(120, 114.8, \frac{13.1}{\sqrt{40}}\right) = 0.006$ There is only a 0.6% chance that the average systolic blood pressure for the randomly selected group is greater than 120.

c. The central limit theorem could not be used if the sample size were four and we did not know the original distribution was normal. The sample size would be too small.

Example 8.4.4

A study was done about violence against prostitutes and the symptoms of the posttraumatic stress that they developed. The age range of the prostitutes was 14 to 61. The mean age was 30.9 years with a standard deviation of nine years.

a. In a sample of 25 prostitutes, what is the probability that the mean age of the prostitutes is less than 35?
b. Is it likely that the mean age of the sample group could be more than 50 years? Interpret the results.
c. In a sample of 49 prostitutes, what is the probability that the sum of the ages is no less than 1,600?
d. Is it likely that the sum of the ages of the 49 prostitutes is at most 1,595? Interpret the results.
e. Find the 95th percentile for the sample mean age of 65 prostitutes. Interpret the results.
f. Find the 90th percentile for the sum of the ages of 65 prostitutes. Interpret the results.

Answer

1. $P(\bar{x} < 35) =$ `normalcdf` $(-E99, 35, 30.9, 1.8) = 0.9886$
2. $P(\bar{x} > 50) =$ `normalcdf` $(50, E99, 30.9, 1.8) \approx 0$ For this sample group, it is almost impossible for the group's average age to be more than 50. However, it is still possible for an individual in this group to have an age greater than 50.
3. $P(\sum x \geq 1,600) =$ `normalcdf` $(1600, E99, 1514.10, 63) = 0.0864$
4. $P(\sum x \leq 1,595) =$ `normalcdf` $(-E99, 1595, 1514.10, 63) = 0.9005$This means that there is a 90% chance that the sum of the ages for the sample group $n = 49$ is at most 1595.
5. The 95th percentile = `invNorm` $(0.95, 30.9, 1.1) = 32.7$This indicates that 95% of the prostitutes in the sample of 65 are younger than 32.7 years, on average.
6. The 90th percentile = `invNorm` $(0.90, 2008.5, 72.56) = 2101.5$This indicates that 90% of the prostitutes in the sample of 65 have a sum of ages less than 2,101.5 years.

Exercise 8.4.4

According to Boeing data, the 757 airliner carries 200 passengers and has doors with a mean height of 72 inches. Assume for a certain population of men we have a mean of 69.0 inches and a standard deviation of 2.8 inches.

a. What mean doorway height would allow 95% of men to enter the aircraft without bending?
b. Assume that half of the 200 passengers are men. What mean doorway height satisfies the condition that there is a 0.95 probability that this height is greater than the mean height of 100 men?
c. For engineers designing the 757, which result is more relevant: the height from part a or part b? Why?

Answer

a. We know that $\mu_x = \mu = 69$ and we have $\sigma_x = 2.8$. The height of the doorway is found to be `invNorm` $(0.95, 69, 2.8) = 73.61$
b. We know that $\mu_x = \mu = 69$ and we have $\sigma_x = 2.8$. So, `invNorm` $(0.95, 69, 0.28) = 69.49$
c. When designing the doorway heights, we need to incorporate as much variability as possible in order to accommodate as many passengers as possible. Therefore, we need to use the result based on part a.

Historical Note: Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the central limit theorem. Binomial probabilities with a small value for n(say, 20) were displayed in a table in a book. To calculate the probabilities with large values of n, you had to use the binomial formula, which could be very complicated. Using the normal approximation to the binomial distribution simplified the process. To compute the normal approximation to the binomial distribution, take a simple random sample from a population. You must meet the conditions for a binomial distribution:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure
- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$); the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x - 0.5$). The number 0.5 is called the continuity correction factor and is used in the following example.

Example 8.4.5

Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

a. Find the probability that **at least 150** favor a charter school.
b. Find the probability that **at most 160** favor a charter school.
c. Find the probability that **more than 155** favor a charter school.
d. Find the probability that **fewer than 147** favor a charter school.
e. Find the probability that **exactly 175** favor a charter school.

Let $X =$ the number that favor a charter school for grades K trough 5. $X \sim B(n, p)$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is X. $Y \sim N(159, 8.6447)$. See The Normal Distribution for help with calculator instructions.

For part a, you **include 150** so $P(X \geq 150)$ has normal approximation $P(Y \geq 149.5) = 0.8641$.

normalcdf $(149.5, 10^{99}, 159, 8.6447) = 0.8641$

For part b, you **include 160** so $P(X \leq 160)$ has normal approximation $P(Y \leq 160.5) = 0.5689$.

normalcdf $(0, 160.5, 159, 8.6447) = 0.5689$

For part c, you **exclude 155** so $P(X > 155)$ has normal approximation $P(y > 155.5) = 0.6572$

normalcdf $(155.5, 10^{99}, 159, 8.6447) = 0.6572$

For part d, you **exclude 147** so $P(X < 147)$ has normal approximation $P(Y < 146.5) = 0.0741$.

normalcdf $(0, 146.5, 159, 8.6447) = 0.0741$

For part e, $P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5) = 0.0083$.

normalcdf $(174.5, 175.5, 159, 8.6447) = 0.0083$

Because of calculators and computer software that let you calculate binomial probabilities for large values of n easily, it is not necessary to use the the normal approximation to the binomial distribution, provided that you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators, and they easily calculate probabilities for the binomial distribution. If you type in "binomial probability distribution calculation" in an Internet browser, you can find at least one online calculator for the binomial.

For Example, the probabilities are calculated using the following binomial distribution: $(n = 300 and p = 0.53)$. Compare the binomial and normal distribution answers. See Discrete Random Variables for help with calculator instructions for the binomial.

$P(X \geq 150)$: 1 - binomialcdf $(300, 0.53, 149) = 0.8641$

$P(X \leq 160)$: binomialcdf $(300, 0.53, 160) = 0.5684$

$P(X > 155)$: 1 - binomialcdf $(300, 0.53, 155) = 0.6576$

$P(X < 147)$: binomialcdf $(300, 0.53, 146) = 0.0742$

$P(X = 175)$:(You use the binomial pdf.) binomialpdf $(300, 0.53, 175) = 0.0083$

Exercise 8.4.5

In a city, 46 percent of the population favor the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.

Answer

0.0401

References

- Data from the Wall Street Journal.
- "National Health and Nutrition Examination Survey." Center for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed May 17, 2013).

Glossary

Exponential Distribution

a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital, notation: $X \sim Exp(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

Mean

a number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation.; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

Uniform Distribution

a continuous random variable (RV) that has equally likely outcomes over the domain, \(a < x < b\); often referred as the **Rectangular Distribution** because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{a+b}{2}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

8.4E: Using the Central Limit Theorem (Exercises)

8.E: Sampling Distributions (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

6.1: The Mean and Standard Deviation of the Sample Mean

Basic

Q6.1.1

Random samples of size 225 are drawn from a population with mean 100 and standard deviation 20. Find the mean and standard deviation of the sample mean.

Q6.1.2

Random samples of size 64 are drawn from a population with mean 32 and standard deviation 5. Find the mean and standard deviation of the sample mean.

Q6.1.3

A population has mean 75 and standard deviation 12.

a. Random samples of size 121 are taken. Find the mean and standard deviation of the sample mean.
b. How would the answers to part (a) change if the size of the samples were 400 instead of 121?

Q6.1.4

A population has mean 5.75 and standard deviation 1.02.

a. Random samples of size 81 are taken. Find the mean and standard deviation of the sample mean.
b. How would the answers to part (a) change if the size of the samples were 25 instead of 81?

Answers

S6.1.1

$\mu_{\bar{X}} = 100,\ \sigma_{\bar{X}} = 1.33$

S6.1.3

a. $\mu_{\bar{X}} = 75,\ \sigma_{\bar{X}} = 1.09$
b. $\mu_{\bar{X}}$ stays the same but $\sigma_{\bar{X}}$ decreases to 0.6

6.2: The Sampling Distribution of the Sample Mean

Basic

1. A population has mean 128 and standard deviation 22.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 36.
   b. Find the probability that the mean of a sample of size 36 will be within 10 units of the population mean, that is, between 118 and 138.
2. A population has mean $1,542$ and standard deviation 246.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 100.
   b. Find the probability that the mean of a sample of size 100 will be within 100 units of the population mean, that is, between $1,442$ and $1,642$.
3. A population has mean 73.5 and standard deviation 2.5.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 30.
   b. Find the probability that the mean of a sample of size 30 will be less than 72.
4. A population has mean 48.4 and standard deviation 6.3.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 64.
   b. Find the probability that the mean of a sample of size 64 will be less than 46.7.

5. A normally distributed population has mean 25.6 and standard deviation 3.3.
   a. Find the probability that a single randomly selected element X of the population exceeds 30.
   b. Find the mean and standard deviation of $\overline{X}$ for samples of size 9.
   c. Find the probability that the mean of a sample of size 9 drawn from this population exceeds 30.
6. A normally distributed population has mean 57.7 and standard deviation 12.1.
   a. Find the probability that a single randomly selected element X of the population is less than 45.
   b. Find the mean and standard deviation of $\overline{X}$ for samples of size 16.
   c. Find the probability that the mean of a sample of size 16 drawn from this population is less than 45.
7. A population has mean 557 and standard deviation 35.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 50.
   b. Find the probability that the mean of a sample of size 50 will be more than 570.
8. A population has mean 16 and standard deviation 1.7.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 80.
   b. Find the probability that the mean of a sample of size 80 will be more than 16.4.
9. A normally distributed population has mean $1,214$ and standard deviation 122.
   a. Find the probability that a single randomly selected element X of the population is between $1,100$ and $1,300$.
   b. Find the mean and standard deviation of $\overline{X}$ for samples of size 25.
   c. Find the probability that the mean of a sample of size 25 drawn from this population is between $1,100$ and $1,300$.
10. A normally distributed population has mean $57,800$ and standard deviation 750.
    a. Find the probability that a single randomly selected element X of the population is between $57,000$ and $58,000$.
    b. Find the mean and standard deviation of $\overline{X}$ for samples of size 100.
    c. Find the probability that the mean of a sample of size 100 drawn from this population is between $57,000$ and $58,000$.
11. A population has mean 72 and standard deviation 6.
    a. Find the mean and standard deviation of $\overline{X}$ for samples of size 45.
    b. Find the probability that the mean of a sample of size 45 will differ from the population mean 72 by at least 2 units, that is, is either less than 70 or more than 74. (Hint: One way to solve the problem is to first find the probability of the complementary event.)
12. A population has mean 12 and standard deviation 1.5.
    a. Find the mean and standard deviation of $\overline{X}$ for samples of size 90.
    b. Find the probability that the mean of a sample of size 90 will differ from the population mean 12 by at least 0.3 unit, that is, is either less than 11.7 or more than 12.3. (Hint: One way to solve the problem is to first find the probability of the complementary event.)

Applications

13. Suppose the mean number of days to germination of a variety of seed is 22, with standard deviation 2.3 days. Find the probability that the mean germination time of a sample of 160 seeds will be within 0.5 day of the population mean.
14. Suppose the mean length of time that a caller is placed on hold when telephoning a customer service center is 23.8 seconds, with standard deviation 4.6 seconds. Find the probability that the mean length of time on hold in a sample of $1,200$ calls will be within 0.5 second of the population mean.
15. Suppose the mean amount of cholesterol in eggs labeled “large” is 186 milligrams, with standard deviation 7 milligrams. Find the probability that the mean amount of cholesterol in a sample of 144 eggs will be within 2 milligrams of the population mean.
16. Suppose that in one region of the country the mean amount of credit card debt per household in households having credit card debt is $\$15,250$, with standard deviation $\$7,125$. Find the probability that the mean amount of credit card debt in a sample of $1,600$ such households will be within $\$300$ of the population mean.
17. Suppose speeds of vehicles on a particular stretch of roadway are normally distributed with mean 36.6 mph and standard deviation 1.7 mph.
    a. Find the probability that the speed X of a randomly selected vehicle is between 35 and 40 mph.
    b. Find the probability that the mean speed $\overline{X}$ of 20 randomly selected vehicles is between 35 and 40 mph.

18. Many sharks enter a state of tonic immobility when inverted. Suppose that in a particular species of sharks the time a shark remains in a state of tonic immobility when inverted is normally distributed with mean 11.2 minutes and standard deviation 1.1 minutes.
    a. If a biologist induces a state of tonic immobility in such a shark in order to study it, find the probability that the shark will remain in this state for between 10 and 13 minutes.
    b. When a biologist wishes to estimate the mean time that such sharks stay immobile by inducing tonic immobility in each of a sample of 12 sharks, find the probability that mean time of immobility in the sample will be between 10 and 13 minutes.
19. Suppose the mean cost across the country of a 30-day supply of a generic drug is $\$46.58$, with standard deviation $\$4.84$. Find the probability that the mean of a sample of 100 prices of 30-day supplies of this drug will be between $\$45$ and $\$50$.
20. Suppose the mean length of time between submission of a state tax return requesting a refund and the issuance of the refund is 47 days, with standard deviation 6 days. Find the probability that in a sample of 50 returns requesting a refund, the mean such time will be more than 50 days.
21. Scores on a common final exam in a large enrollment, multiple-section freshman course are normally distributed with mean 72.7 and standard deviation 13.1.
    a. Find the probability that the score X on a randomly selected exam paper is between 70 and 80.
    b. Find the probability that the mean score $\overline{X}$ of 38 randomly selected exam papers is between 70 and 80.
22. Suppose the mean weight of school children's bookbags is 17.4 pounds, with standard deviation 2.2 pounds. Find the probability that the mean weight of a sample of 30 bookbags will exceed 17 pounds.
23. Suppose that in a certain region of the country the mean duration of first marriages that end in divorce is 7.8 years, standard deviation 1.2 years. Find the probability that in a sample of 75 divorces, the mean age of the marriages is at most 8 years.
24. Borachio eats at the same fast food restaurant every day. Suppose the time X between the moment Borachio enters the restaurant and the moment he is served his food is normally distributed with mean 4.2 minutes and standard deviation 1.3 minutes.
    a. Find the probability that when he enters the restaurant today it will be at least 5 minutes until he is served.
    b. Find the probability that average time until he is served in eight randomly selected visits to the restaurant will be at least 5 minutes.

Additional Exercises

25. A high-speed packing machine can be set to deliver between 11 and 13 ounces of a liquid. For any delivery setting in this range the amount delivered is normally distributed with mean some amount μ and with standard deviation 0.08 ounce. To calibrate the machine it is set to deliver a particular amount, many containers are filled, and 25 containers are randomly selected and the amount they contain is measured. Find the probability that the sample mean will be within 0.05 ounce of the actual mean amount being delivered to all containers.
26. A tire manufacturer states that a certain type of tire has a mean lifetime of $60,000$ miles. Suppose lifetimes are normally distributed with standard deviation $\sigma = 3,500$ miles.
    a. Find the probability that if you buy one such tire, it will last only $57,000$ or fewer miles. If you had this experience, is it particularly strong evidence that the tire is not as good as claimed?
    b. A consumer group buys five such tires and tests them. Find the probability that average lifetime of the five tires will be $57,000$ miles or less. If the mean is so low, is that particularly strong evidence that the tire is not as good as claimed?

Answers

1. a. $\mu_{\overline{X}} = 128,\ \sigma_{\overline{X}} = 3.67$
   b. 0.9936
2.
3. a. $\mu_{\overline{X}} = 73.5,\ \sigma_{\overline{X}} = 0.456$
   b. 0.0005
4.
5. a. 0.0918
   b. $\mu_{\overline{X}} = 25.6,\ \sigma_{\overline{X}} = 1.1$
   c. 0.0000
6.

7. a. $\mu_{\overline{X}} = 557,\ \sigma_{\overline{X}} = 4.9497$
   b. 0.0043
8.
9. a. 0.5818
   b. $\mu_{\overline{X}} = 1214\ \sigma_{\overline{X}} = 24.4$
   c. 0.9998
10.
11. a. $\mu_{\overline{X}} = 72\ \sigma_{\overline{X}} = 0.8944$
    b. 0.0250
12.
13. 0.9940
14.
15. 0.9994
16.
17. a. 0.8036
    b. 1.0000
18.
19. 0.9994
20.
21. a. 0.2955
    b. 0.8977
22.
23. 0.9251
24.
25. 0.9982

6.3: The Sample Proportion

Basic

1. The proportion of a population with a characteristic of interest is $p = 0.37$. Find the mean and standard deviation of the sample proportion $\widehat{P}$ obtained from random samples of size $1,600$.
2. The proportion of a population with a characteristic of interest is $p = 0.82$. Find the mean and standard deviation of the sample proportion $\widehat{P}$ obtained from random samples of size 900.
3. The proportion of a population with a characteristic of interest is $p = 0.76$. Find the mean and standard deviation of the sample proportion $\widehat{P}$ obtained from random samples of size $1,200$.
4. The proportion of a population with a characteristic of interest is $p = 0.37$. Find the mean and standard deviation of the sample proportion $\widehat{P}$ obtained from random samples of size 125.
5. Random samples of size 225 are drawn from a population in which the proportion with the characteristic of interest is 0.25. Decide whether or not the sample size is large enough to assume that the sample proportion $\widehat{P}$ is normally distributed.
6. Random samples of size $1,600$ are drawn from a population in which the proportion with the characteristic of interest is 0.05. Decide whether or not the sample size is large enough to assume that the sample proportion $\widehat{P}$ is normally distributed.
7. Random samples of size n produced sample proportions $\hat{p}$ as shown. In each case decide whether or not the sample size is large enough to assume that the sample proportion $\widehat{P}$ is normally distributed.
   a. $n = 50,\ \hat{p} = 0.48$
   b. $n = 50,\ \hat{p} = 0.12$
   c. $n = 100,\ \hat{p} = 0.12$
8. Samples of size n produced sample proportions $\hat{p}$ as shown. In each case decide whether or not the sample size is large enough to assume that the sample proportion $\widehat{P}$ is normally distributed.
   a. $n = 30,\ \hat{p} = 0.72$

b. $n = 30,\ \hat{p} = 0.84$
c. $n = 75,\ \hat{p} = 0.84$

9. A random sample of size 121 is taken from a population in which the proportion with the characteristic of interest is $p = 0.47$. Find the indicated probabilities.
   a. $P(0.45 \le \widehat{P} \le 0.50)$
   b. $P(\widehat{P} \ge 0.50)$
10. A random sample of size 225 is taken from a population in which the proportion with the characteristic of interest is $p = 0.34$. Find the indicated probabilities.
   a. $P(0.25 \le \widehat{P} \le 0.40)$
   b. $P(\widehat{P} \ge 0.35)$
11. A random sample of size 900 is taken from a population in which the proportion with the characteristic of interest is p = 0.62. Find the indicated probabilities.
   a. $P(0.60 \le \widehat{P} \le 0.64)$
   b. $P(0.57 \le \widehat{P} \le 0.67)$
12. A random sample of size 1,100 is taken from a population in which the proportion with the characteristic of interest is p = 0.28. Find the indicated probabilities.
   1. $P(0.27 \le \widehat{P} \le 0.29)$
   2. $P(0.23 \le \widehat{P} \le 0.33)$

Applications

13. Suppose that 8% of all males suffer some form of color blindness. Find the probability that in a random sample of 250 men at least 10% will suffer some form of color blindness. First verify that the sample is sufficiently large to use the normal distribution.
14. Suppose that 29% of all residents of a community favor annexation by a nearby municipality. Find the probability that in a random sample of 50 residents at least 35% will favor annexation. First verify that the sample is sufficiently large to use the normal distribution.
15. Suppose that 2% of all cell phone connections by a certain provider are dropped. Find the probability that in a random sample of $1,500$ calls at most 40 will be dropped. First verify that the sample is sufficiently large to use the normal distribution.
16. Suppose that in 20% of all traffic accidents involving an injury, driver distraction in some form (for example, changing a radio station or texting) is a factor. Find the probability that in a random sample of 275 such accidents between 15% and 25% involve driver distraction in some form. First verify that the sample is sufficiently large to use the normal distribution.
17. An airline claims that 72% of all its flights to a certain region arrive on time. In a random sample of 30 recent arrivals, 19 were on time. You may assume that the normal distribution applies.
   a. Compute the sample proportion.
   b. Assuming the airline's claim is true, find the probability of a sample of size 30 producing a sample proportion so low as was observed in this sample.
18. A humane society reports that 19% of all pet dogs were adopted from an animal shelter. Assuming the truth of this assertion, find the probability that in a random sample of 80 pet dogs, between 15% and 20% were adopted from a shelter. You may assume that the normal distribution applies.
19. In one study it was found that 86% of all homes have a functional smoke detector. Suppose this proportion is valid for all homes. Find the probability that in a random sample of 600 homes, between 80% and 90% will have a functional smoke detector. You may assume that the normal distribution applies.
20. A state insurance commission estimates that 13% of all motorists in its state are uninsured. Suppose this proportion is valid. Find the probability that in a random sample of 50 motorists, at least 5 will be uninsured. You may assume that the normal distribution applies.
21. An outside financial auditor has observed that about 4% of all documents he examines contain an error of some sort. Assuming this proportion to be accurate, find the probability that a random sample of 700 documents will contain at least 30 with some sort of error. You may assume that the normal distribution applies.
22. Suppose 7% of all households have no home telephone but depend completely on cell phones. Find the probability that in a random sample of 450 households, between 25 and 35 will have no home telephone. You may assume that the normal

distribution applies.

Additional Exercises

23. Some countries allow individual packages of prepackaged goods to weigh less than what is stated on the package, subject to certain conditions, such as the average of all packages being the stated weight or greater. Suppose that one requirement is that at most 4% of all packages marked 500 grams can weigh less than 490 grams. Assuming that a product actually meets this requirement, find the probability that in a random sample of 150 such packages the proportion weighing less than 490 grams is at least 3%. You may assume that the normal distribution applies.
24. An economist wishes to investigate whether people are keeping cars longer now than in the past. He knows that five years ago, 38% of all passenger vehicles in operation were at least ten years old. He commissions a study in which 325 automobiles are randomly sampled. Of them, 132 are ten years old or older.
    a. Find the sample proportion.
    b. Find the probability that, when a sample of size 325 is drawn from a population in which the true proportion is 0.38, the sample proportion will be as large as the value you computed in part (a). You may assume that the normal distribution applies.
    c. Give an interpretation of the result in part (b). Is there strong evidence that people are keeping their cars longer than was the case five years ago?
25. A state public health department wishes to investigate the effectiveness of a campaign against smoking. Historically 22% of all adults in the state regularly smoked cigars or cigarettes. In a survey commissioned by the public health department, 279 of $1,500$ randomly selected adults stated that they smoke regularly.
    a. Find the sample proportion.
    b. Find the probability that, when a sample of size $1,500$ is drawn from a population in which the true proportion is 0.22, the sample proportion will be no larger than the value you computed in part (a). You may assume that the normal distribution applies.
    c. Give an interpretation of the result in part (b). How strong is the evidence that the campaign to reduce smoking has been effective?
26. In an effort to reduce the population of unwanted cats and dogs, a group of veterinarians set up a low-cost spay/neuter clinic. At the inception of the clinic a survey of pet owners indicated that 78% of all pet dogs and cats in the community were spayed or neutered. After the low-cost clinic had been in operation for three years, that figure had risen to 86%.
    a. What information is missing that you would need to compute the probability that a sample drawn from a population in which the proportion is 78% (corresponding to the assumption that the low-cost clinic had had no effect) is as high as 86%?
    b. Knowing that the size of the original sample three years ago was 150 and that the size of the recent sample was 125, compute the probability mentioned in part (a). You may assume that the normal distribution applies.
    c. Give an interpretation of the result in part (b). How strong is the evidence that the presence of the low-cost clinic has increased the proportion of pet dogs and cats that have been spayed or neutered?
27. An ordinary die is "fair" or "balanced" if each face has an equal chance of landing on top when the die is rolled. Thus the proportion of times a three is observed in a large number of tosses is expected to be close to $1/6$ or $0.1\bar{6}$. Suppose a die is rolled 240 times and shows three on top 36 times, for a sample proportion of 0.15.
    a. Find the probability that a fair die would produce a proportion of 0.15 or less. You may assume that the normal distribution applies.
    b. Give an interpretation of the result in part (b). How strong is the evidence that the die is not fair?
    c. Suppose the sample proportion 0.15 came from rolling the die $2,400$ times instead of only 240 times. Rework part (a) under these circumstances.
    d. Give an interpretation of the result in part (c). How strong is the evidence that the die is not fair?

Answers

1. $\mu_{\widehat{P}} = 0.37,\ \sigma_{\widehat{P}} = 0.012$

3. $\mu_{\widehat{P}} = 0.76,\ \sigma_{\widehat{P}} = 0.012$

5. $p \pm 3\sqrt{\frac{pq}{n}} = 0.25 \pm 0.087,$ yes

7. a. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.48 \pm 0.21$, yes
   b. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.12 \pm 0.14$, no
   c. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.12 \pm 0.10$, yes

9. a. 0.4154
   b. 0.2546

11. a. 0.7850
    b. 0.9980

13. $p \pm 3\sqrt{\frac{pq}{n}} = 0.08 \pm 0.05$ and $[0.03, 0.13] \subset [0, 1], 0.1210$

15. $p \pm 3\sqrt{\frac{pq}{n}} = 0.02 \pm 0.01$ and $[0.01, 0.03] \subset [0, 1], 0.9671$

17. a. 0.63
    b. 0.1446

19. 0.9977

21. 0.3483

23. 0.7357

25. a. 0.186
    b.0.0007

In a population in which the true proportion is 22% the chance that a random sample of size $1,500$ would produce a sample proportion of 18.6% or less is only $7/100$ of 1%. This is strong evidence that currently a smaller proportion than 22% smoke.

27. a. 0.2451

We would expect a sample proportion of 0.15 or less in about 24.5% of all samples of size 240, so this is practically no evidence at all that the die is not fair.

   b. 0.0139

We would expect a sample proportion of 0.15 or less in only about 1.4% of all samples of size $2,400$, so this is strong evidence that the die is not fair.

Contributor

- Anonymous

CHAPTER OVERVIEW

9: Confidence Intervals

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

9.1: Prelude to Confidence Intervals
9.2: A Single Population Mean using the Normal Distribution
9.3: A Single Population Mean using the Student t-Distribution
9.4: A Population Proportion
9.5: Confidence Interval - Home Costs (Worksheet)
9.6: Confidence Interval -Place of Birth (Worksheet)
9.7: Confidence Interval -Women's Heights (Worksheet)
9.E: Confidence Intervals (Exercises)
9.S: Confidence Intervals (Summary)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

9.1: Prelude to Confidence Intervals

Learning Objectives

By the end of this chapter, the student should be able to:

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
- Interpret the Student's t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the Student's *t* distributions.
- Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals.

Figure 9.1.1. Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the sample standard deviation, s. You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the point estimate for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

Each of $\bar{x}$ and s is called a statistic.

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose, for the iTunes example, we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1. \tag{9.1.1}$$

The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. For our iTunes example, two standard deviations is (2)(0.1) = 0.2. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ.

Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \tag{9.1.2}$$

and

$$\bar{x} + 0.2 = 2 + 0.2 = 2.2 \tag{9.1.3}$$

We say that we are 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The 95% confidence interval is (1.8, 2.2). This 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, $\bar{x}$. Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

> Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

Collaborative Exercise

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. Let $\sigma = 3$ and n = the number of students surveyed.
3. Construct the interval $\left(\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}}\right)$.

We say we are approximately 95% confident that the true mean number of meals that students eat out in a week is between _____ and _____.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation),
- the sample and its size.

Inferential Statistics

also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of

the production is defective.

Parameter

a numerical characteristic of a population

Point Estimate

a single number computed from a sample and used to estimate a population parameter

---

9.2: A Single Population Mean using the Normal Distribution

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ, where the population standard deviation is known, we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (EBM) is called the error bound for a population mean (abbreviated *EBM*). The sample mean $\bar{x}$ is the point estimate of the unknown population mean μ.

The confidence interval estimate will have the form:

$$(\text{point estimate} - \text{error bound}, \text{point estimate} + \text{error bound})$$

or, in symbols,

$$(\bar{x} - EBM, \bar{x} + EBM)$$

The **margin of error** (EBM) depends on the confidence level (abbreviated CL). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level, CL. α is the probability that the interval does not contain the unknown population parameter. Mathematically,

$$\alpha + CL = 1.$$

Example 9.2.1

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population. The sample mean is seven, and the error bound for the mean is 2.5: $\bar{x} = 7$ and $EBM = 2.5$

The confidence interval is (7 – 2.5, 7 + 2.5) and calculating the values gives (4.5, 9.5). If the confidence level (CL) is 95%, then we say that, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

Exercise 9.2.1

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2. What is the confidence interval estimate for the population mean?

Answer

(11.8, 18.2)

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$. To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10$ in both tails, or 5% in each tail, of the normal distribution.

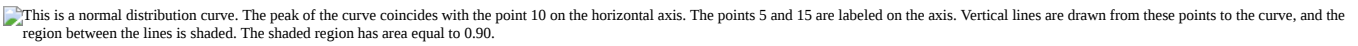

This is a normal distribution curve. The peak of the curve coincides with the point 10 on the horizontal axis. The points 5 and 15 are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 0.90.

Figure 9.2.1

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the *z*-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is

$$\frac{\sigma}{\sqrt{n}}$$

This fraction is commonly called the "standard error of the mean" to distinguish clearly the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the central limit theorem:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N(\mu_x, \frac{\sigma}{\sqrt{n}})$.
- When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound EBM.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95, \alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

using a calculator, computer or a standard normal probability table.

`invNorm` $(0.975, 0, 1) = 1.96$

Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution $Z \sim N(0, 1)$.

Calculating the Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

$$EBM = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Constructing the Confidence Interval

The confidence interval estimate has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$$

Figure 8.2.2.

Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example 9.2.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Answer

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\sigma = 3; n = 36$$

The confidence level is 90% (CL = 0.90)

$$CL = 0.90$$

so

$$\alpha = 1 - CL = 1 - 0.90 = 0.10$$

$$\frac{\alpha}{2} = 0.05 z_{\frac{\alpha}{2}} = z_{0.05}$$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is $1 - 0.05 = 0.95$.

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

using $\text{invNorm}(0.95, 0, 1)$ on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$$

$$\bar{x} - EBM = 68 - 0.8225 = 67.1775$$

$$\bar{x} + EBM = 68 + 0.8225 = 68.8225$$

The 90% confidence interval is **(67.1775, 68.8225).**

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `7:ZInterval` .
Press `ENTER` .
Arrow to `Stats` and press `ENTER` .
Arrow down and enter three for σ, 68 for $\bar{x}$, 36 for n, and .90 for `C-level` .
Arrow down to `Calculate` and press `ENTER` .
The confidence interval is (to three decimal places)(67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Exercise 9.2.2

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.1347, 37.8653)

Example 9.2.3: Specific Absorption Rate

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. Table shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

Solution A

To find the confidence interval, start by finding the point estimate: the sample mean.

$$\bar{x} = 1.024$$

Next, find the EBM. Because you are creating a 98% confidence interval, $CL = 0.98$.

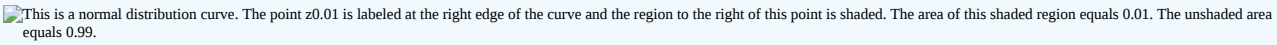

This is a normal distribution curve. The point z0.01 is labeled at the right edge of the curve and the region to the right of this point is shaded. The area of this shaded region equals 0.01. The unshaded area equals 0.99.

Figure 8.2.3.

You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\bar{x} \pm EBM$.

$\bar{x} - EBM = 1.024 - 0.1431 = 0.8809$

$\bar{x} - EBM = 1.024 - 0.1431 = 1.1671$

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution B

- Press STAT and arrow over to TESTS.
- Arrow down to 7:Z Interval.
- Press ENTER.
- Arrow to Stats and press ENTER.
- Arrow down and enter the following values:
  - σ : 0.337
  - $\bar{x}$: 1024
  - n : 30
  - C-level: 0.98
- Arrow down to Calculate and press ENTER.
- The confidence interval is (to three decimal places) (0.881, 1.167).

Exercise 9.2.3

Table shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Answer

$$\bar{x} = 0.940$$

$$\frac{\alpha}{2} = \frac{1 - CL}{2} = \frac{1 - 0.93}{2} = 0.035$$

$$z_{0.035} = 1.812$$

$$EBM = (z_{0.035})\left(\frac{\sigma}{\sqrt{n}}\right) = (1.812)\left(\frac{0.337}{\sqrt{20}}\right) = 0.1365$$

$$\bar{x} - EBM = 0.940 - 0.1365 = 0.8035$$

$$\bar{x} + EBM = 0.940 + 0.1365 = 1.0765$$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Notice the difference in the confidence intervals calculated in Example and the following Try It exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example 9.2.4

Suppose we change the original problem in Example by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Answer

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = \left(z\frac{\alpha}{2}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$\sigma = 3; n = 36$; The confidence level is 95% (CL = 0.95).

$$CL = 0.95 \text{ so } \alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 z\frac{\alpha}{2} = z_{0.025}$$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$

$$z\frac{\alpha}{2} = z_{0.025} = 1.96$$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$$

$$\bar{x} - EBM = 68 - 0.98 = 67.02$$

$$\bar{x} + EBM = 68 + 0.98 = 68.98$$

Notice that the EBM is larger for a 95% confidence level in the original problem.

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.

Figure 8.2.4.

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

? Exercise 9.2.4

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Answer

(33.37, 38.63)

✔ Example 9.2.5

Suppose we change the original problem in Example to see what happens to the error bound if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease

the sample size to $n = 25$ instead of $n = 36$?

- $\bar{x} = 68$
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; The confidence level is 90% (CL=0.90); $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Answer

Solution A

If we **increase** the sample size n to 100, we **decrease** the error bound.

When $n = 100$: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935.$

Solution B

If we **decrease** the sample size n to 25, we **increase** the error bound.

When $n = 25$: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987.$

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

? Exercise 9.2.5

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.6041, 37.3958)

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example 9.2.6

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68 : $EBM = 68.82 - 68 = 0.82$.
- If we don't know the sample mean: $EBM = \dfrac{(68.82 - 67.18)}{2} = 0.82$.

Calculate the Sample Mean:

- If we know the error bound: $\bar{x} = 68.82 - 0.82 = 68$
- If we don't know the error bound: $\bar{x} = \dfrac{(67.18 + 68.82)}{2} = 68$.

Exercise 9.2.6

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Answer

Sample mean is 45, error bound is 2.88

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population mean when the population standard deviation is known is

$$EBM = \left(z\frac{a}{2}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

The formula for sample size is $n = \dfrac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n. In Equation ???, z is $z\frac{a}{2}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 9.2.7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

Solution

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.
- $n = \dfrac{z^2\sigma^2}{EBM^2} = \dfrac{(1.96)^2(15)^2}{2^2}$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Exercise 9.2.7

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Answer

35 students

References

1. “American Fact Finder.” U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...html?refresh=t (accessed July 2, 2013).
2. “Disclosure Data Catalog: Candidate Summary Report 2012.” U.S. Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2, 2013).
3. “Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall.” Foothill De Anza Community College District. Available online at research.fhda.edu/factbook/FH...phicTrends.htm (accessed September 30,2013).
4. Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. “2000 CDC Growth Charts for the United States: Methods and Development.” Centers for Disease Control and Prevention. Available online at www.cdc.gov/growthcharts/2000...thchart-us.pdf (accessed July 2, 2013).
5. La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels/ (accessed July 2, 2013).
6. “Mean Income in the Past 12 Months (in 2011 Inflaction-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates.” American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...prodType=table (accessed July 2, 2013).
7. “Metadata Description of Candidate Summary File.” U.S. Federal Election Commission. Available online at www.fec.gov/finance/disclosur...esummary.shtml (accessed July 2, 2013).
8. “National Health and Nutrition Examination Survey.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed July 2, 2013).

Glossary

Confidence Level (CL)

the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the $CL = 90$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Error Bound for a Population Mean (EBM)

the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

9.3: A Single Population Mean using the Student t-Distribution

In practice, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the Student's t-distribution. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever s is used as an estimate for σ. If you draw a simple random sample of size n from a population that has an approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}, \tag{9.3.1}$$

then the t-scores follow a Student's t-distribution with $n-1$ degrees of freedom. The t-score has the same interpretation as the z-score. It measures how far $\bar{x}$ is from its mean μ. For each sample size n, there is a different Student's t-distribution.

The degrees of freedom, $n-1$, come from the calculation of the sample standard deviation s. Previously, we used n deviations ($x - \bar{x}$ values) to calculate s. Because the sum of the deviations is zero, we can find the last deviation once we know the other $n-1$ deviations. The other $n-1$ deviations can change or vary freely. We call the number $n-1$ the degrees of freedom (df).

> *For each sample size n, there is a different Student's t-distribution.*

Properties of the Student's t-Distribution

- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's t-probabilities. The TI-83,83+, and 84+ have a tcdf function to find the probability for given values of t. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's t-distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator,

you need to use a probability table for the Student's t-Distribution.) When using a t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's t-table gives t-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's t-probabilities.**

The notation for the Student's t-distribution (using *T* as the random variable) is:

- $T \sim t_{df}$ where $df = n - 1$.
- For example, if we have a sample of size $n = 20$ items, then we calculate the degrees of freedom as $df = n - 1 = 20 - 1 = 19$ and we write the distribution as $T \sim t_{19}$.

If the population standard deviation is not known, the error bound for a population mean is:

- $EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n - 1$ degrees of freedom, and
- $s =$ sample standard deviation.

The format for the confidence interval is:

$$(\bar{x} - EBM, \bar{x} + EBM). \tag{9.3.2}$$

To calculate the confidence interval directly:

Press STAT.
Arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or just press 8).

Example 9.3.1: Acupuncture

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 8.2267$$

$$s = 1.6722 \; n = 15$$

$$df = 15 - 1 = 14 CL so \alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 t_{\frac{\alpha}{2}} = t_{0.025}$$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is 1 – 0.025 = 0.975

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$ using invT(.975,14) on the TI-84+ calculator.

$$EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

$$= (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

Now it is just a direct application of Equation 9.3.2:

$$\bar{x} - EBM = 8.2267 - 0.9240 = 7.3$$
$$\bar{x} + EBM = 8.2267 + 0.9240 = 9.15$$

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`).
Arrow to `Data` and press `ENTER` .
Arrow down to `List` and enter the list name where you put the data.
There should be a 1 after `Freq` .
Arrow down to `C-level` and enter 0.95
Arrow down to `Calculate` and press `ENTER` .
The 95% confidence interval is (7.3006, 9.1527)

When calculating the error bound, a probability table for the Student's t-distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

Exercise 9.3.1

You do a study of hypnotherapy to determine how effective it is in increasing the number of hourse of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Answer

(8.1634, 9.8032)

Example 9.3.2: The Human Toxome Project

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. Table 9.3.1 shows how many of the targeted chemicals were found in each infant's cord blood.

Table 9.3.1

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an in infant's blood.

Solution A

From the sample, you can calculate $\bar{x} = 127.45$ and $s = 25.965$. There are 20 infants in the sample, so $n = 20$, and $df = 20 - 1 = 19$.

You are asked to calculate a 90% confidence interval: $CL = 0.90$, so

$$\alpha = 1 - CL = 1 - 0.90 = 0.10 \frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05} \tag{9.3.3}$$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is $1 - 0.05 = 0.95$.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 1.729 \left(\frac{25.965}{\sqrt{20}} \right) \approx 10.038$$

$$\bar{x} - EBM = 127.45 - 10.038 = 117.412$$

$$\bar{x} + EBM = 127.45 + 10.038 = 137.488$$

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution B

Enter the data as a list.

Press `STAT` and arrow over to `TESTS` .
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`). Arrow to Data and press `ENTER` .
Arrow down to `List` and enter the list name where you put the data.
Arrow down to `Freq` and enter 1.
Arrow down to `C-level` and enter 0.90
Arrow down to `Calculate` and press `ENTER` .

The 90% confidence interval is (117.41, 137.49).

Example 9.3.3

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in Table 9.3.2. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

Table 9.3.2

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Solution A

- $\bar{x} = 6.133$,
- $s = 5.514$,
- $n = 15$, and
- $df = 15 - 1 = 14$.

$$CL = 0.98, \text{ so } \alpha = 1 - CL = 1 - 0.98 = 0.02$$

$$\frac{\alpha}{2} = 0.01 t_{\frac{\alpha}{2}} = t_{0.01} 2.624$$

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 2.624 \left(\frac{5.514}{\sqrt{15}} \right) - 3.736$$

$$\bar{x} - EBM = 6.133 - 3.736 = 2.397$$

$$\bar{x} + EBM = 6.133 + 3.736 = 9.869$$

We estimate with 98% confidence that the mean number of all hours that statistics students spend watching television in one week is between 2.397 and 9.869.

Solution B

Enter the data as a list.

Press STAT and arrow over to TESTS .
Arrow down to 8:TInterval .
Press ENTER .
Arrow to Data and press ENTER .
Arrow down and enter the name of the list where the data is stored.
Enter Freq : 1
Enter C-Level : 0.98
Arrow down to Calculate and press Enter .
The 98% confidence interval is (2.3965, 9,8702).

Reference

1. “America’s Best Small Companies.” Forbes, 2013. Available online at http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).
2. Data from *Microsoft Bookshelf*.
3. Data from http://www.businessweek.com/.
4. Data from http://www.forbes.com/.
5. “Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012.” Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2,2013).
6. “Human Toxome Project: Mapping the Pollution in People.” Environmental Working Group. Available online at www.ewg.org/sites/humantoxome...tero%2Fnewborn (accessed July 2, 2013).
7. “Metadata Description of Leadership PAC List.” Federal Election Commission. Available online at www.fec.gov/finance/disclosur...pPacList.shtml (accessed July 2, 2013).

Glossary

Degrees of Freedom (df)

the number of objects in a sample that are free to vary

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation

Student's t-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as *n* get larger.
- There is a "family" of t–distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

9.4: A Population Proportion

During an election year, we see articles in the newspaper that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 – 0.03,0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the error bound, and the confidence level for a proportion is similar to that for the population mean, but the formulas are different. How do you know you are dealing with a proportion problem? First, the underlying distribution is a binomial distribution. (There is no mention of a mean or average.) If X is a binomial random variable, then

$$X \sim B(n, p)$$

where n is the number of trials and p is the probability of a success.

To form a proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable P' (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\hat{P}$, read "P hat".)

When n is large and p is not close to zero or one, we can use the normal distribution to approximate the binomial.

$$X \sim N(np, \sqrt{npq})$$

If we divide the random variable, the mean, and the standard deviation by n, we get a normal distribution of proportions with P', called the estimated proportion, as the random variable. (Recall that a proportion as the number of successes divided by n.)

$$\frac{X}{n} = P' \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$

Using algebra to simplify:

$$\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

P' follows a normal distribution for proportions:

$$\frac{X}{n} = P' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

The confidence interval has the form

$$(p' - EBP, p' + EBP).$$

where

- EBP is error bound for the proportion.
- $p' = \frac{x}{n}$
- $p' =$ the estimated proportion of successes (p' is a point estimate for p, the true proportion.)
- $x =$ the number of successes
- $n =$ the size of the sample

The error bound (EBP) for a proportion is

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right)$$

where $q = 1 - p'$.

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is

$$\sqrt{\frac{pq}{n}}.$$

However, in the error bound formula, we use

$$\sqrt{\frac{p'q'}{n}}$$

as the standard deviation, instead of

$$\sqrt{\frac{pq}{n}}.$$

In the error bound formula, the sample proportions p' and q' are estimates of the unknown population proportions p and q. The estimated proportions p' and q' are used because p and q are not known. The sample proportions p' and q' are calculated from the data: p' is the estimated proportion of successes, and q' is the estimated proportion of failures.

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five.

Normal Distribution of Proportions

For the normal distribution of proportions, the z-score formula is as follows.

If

$$P' \sim N\left(p, \sqrt{\frac{pq}{n}}\right) \tag{9.4.1}$$

then the z-score formula is

$$z = \frac{p' - p}{\sqrt{\frac{pq}{n}}} \tag{9.4.2}$$

Example 9.4.1

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution A

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Let $X =$ the number of people in the sample who have cell phones. X is binomial.

$$X \sim B(500, \frac{421}{500}).$$

To calculate the confidence interval, you must find p', q', and EBP.

- $n = 500$
- $x =$ the number of successes $= 421$

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

- $p' = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$$q' = 1 - p' = 1 - 0.842 = 0.158$$

Since $CL = 0.95$, then

$$\alpha = 1 - CL = 1 - 0.95 = 0.05 \left(\frac{\alpha}{2}\right) = 0.025.$$

Then

$$z\frac{\alpha}{2} = z_{0.025 = 1.96}$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.975,0,1) to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = \left(z\frac{\alpha}{2}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$p' - EBP = 0.842 - 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.810, 0.874)$.

Interpretation

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `A:1-PropZint` . Press `ENTER` .
Arrow down to xx and enter 421.
Arrow down to nn and enter 500.
Arrow down to `C-Level` and enter .95.
Arrow down to `Calculate` and press `ENTER` .
The confidence interval is (0.81003, 0.87397).

? Exercise 9.4.1

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Answer

(0.3315, 0.4525)

Example 9.4.2

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A

- $x = 300$ and
- $n = 500$

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

Since $CL = 0.90$, then

$$\alpha = 1 - CL = 1 - 0.90 = 0.10 \left(\frac{\alpha}{2}\right) = 0.05 \tag{9.4.3}$$

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.95,0,1) to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.645)\sqrt{\frac{(0.60)(0.40)}{500}} = 0.036$$

$$p' - EBP = 0.60 - 0.036 = 0.564$$

$$p' + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.564, 0.636)$.

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `A:1-PropZint` . Press `ENTER` .
Arrow down to xx and enter 300.
Arrow down to nn and enter 500.
Arrow down to `C-Level` and enter 0.90.
Arrow down to `Calculate` and press `ENTER` .

The confidence interval is (0.564, 0.636).

Exercise 9.4.2

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.
b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

Answer a

(0.7731, 0.8269); We estimate with 90% confidence that the true percent of all students in the district who are against the new legislation is between 77.31% and 82.69%.

Answer b

Sixty-eight percent (68%) of students own an iPod and a smart phone.

$$p' = 0.68$$

$$q' = 1 - p' = 1 - 0.68 = 0.32$$

Since $CL = 0.97$, we know

$$\alpha = 1 - 0.97 = 0.03$$

and

$$\frac{\alpha}{2} = 0.015.$$

The area to the left of $z_{0.05}$ is 0.015, and the area to the right of $z_{0.05}$ is 1 – 0.015 = 0.985.

Using the TI 83, 83+, or 84+ calculator function InvNorm(0.985,0,1),

$$z_{0.05} = 2.17$$

$$EPB = \left(z\frac{\alpha}{2}\right)\sqrt{\frac{p'q'}{n}} = 2.17\sqrt{\frac{0.68(0.32)}{300}} \approx 0.0269$$

$$p' - EPB = 0.68 - 0.0269 = 0.6531$$

$$p' + EPB = 0.68 + 0.0269 = 0.7069$$

We are 97% confident that the true proportion of all students who own an iPod and a smart phone is between 0.6531 and 0.7069.

Calculator

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 300*0.68.
Arrow down to n and enter 300.
Arrow down to C-Level and enter 0.97.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.6531, 0.7069).

"Plus Four" Confidence Interval for p

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is $n+4$, and the new count of successes is $x+2$. Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example 9.4.3

A random sample of 25 statistics students was asked: "Have you smoked a cigarette in the past week?" Six students reported smoking within the past week. Use the "plus-four" method to find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution A

Six students out of 25 reported smoking within the past week, so $x=6$ and $n=25$. Because we are using the "plus-four" method, we will use $x=6+2=8$ and $n=25+4=29$.

$$p'=\frac{x}{n}=\frac{8}{29}\approx 0.276$$

$$q'=1-p'=1-0.276=0.724$$

Since $CL=0.95$, we know $\alpha=1-0.95=0.05$ and $\frac{\alpha}{2}=0.025$.

$$z_{0.025}=1.96$$

$$EPB=\left(z\frac{\alpha}{2}\right)\sqrt{\frac{p'q'}{n}}=(1.96)\sqrt{\frac{0.276(0.724)}{29}}\approx 0.163$$

$$p'-EPB=0.276-0.163=0.113$$

$$p'+EPB=0.276+0.163=0.439$$

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 0.113 and 0.439.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

REMINDER

Remember that the plus-four method assume an additional four trials: two successes and two failures. You do not need to change the process for calculating the confidence interval; simply update the values of x and n to reflect these additional trials.

Arrow down to x and enter eight.

Arrow down to n and enter 29.
Arrow down to C-Level and enter 0.95.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.113, 0.439).

Exercise 9.4.3

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the “plus-four” method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Solution A

Using “plus four,” we have $x = 31 + 2 = 33$ and $n = 65 + 4 = 69$.

$$p' = \frac{33}{69} \approx 0.478$$

$$q' = 1 - p' = 1 - 0.478 = 0.522$$

Since $CL = 0.96$, we know $\alpha = 1 - 0.96 = 0.04$ and $\frac{\alpha}{2} = 0.02$.

$$z_{0.02} = 2.054$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (2.054)\left(\sqrt{\frac{(0.478)(0.522)}{69}}\right) = 0.124$$

$$p' - EPB = 0.478 - 0.124 = 0.354$$

$$p' + EPB = 0.478 + 0.124 = 0.602$$

We are 96% confident that between 35.4% and 60.2% of all freshmen at State U have declared a major.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 33.
Arrow down to n and enter 69.
Arrow down to C-Level and enter 0.96.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.355, 0.602).

Example 9.4.4

The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the “plus four” method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

Solution A

Using “plus-four,” we have $x = 13 + 2 = 15$ and $n = 50 + 4 = 54$.

$$p' = \frac{15}{54} \approx 0.278$$

$$q' = 1 - p' = 1 - 0.241 = 0.722$$

Since $CL = 0.90$, we know $\alpha = 1 - 0.90 = 0.10$ and $\frac{\alpha}{2} = 0.05$.

$$z_{0.05} = 1.645$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.278)(0.722)}{54}}\right) \approx 0.100$$

$$p' - EPB = 0.278 - 0.100 = 0.178$$

$$p' + EPB = 0.278 + 0.100 = 0.378$$

We are 90% confident that between 17.8% and 37.8% of all teens would report having more than 500 friends on Facebook.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 15.
Arrow down to n and enter 54.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.178, 0.378).

Exercise 9.4.4

The Berkman Center Study referenced in Example talked to teens in smaller focus groups, but also interviewed additional teens over the phone. When the study was complete, 588 teens had answered the question about their Facebook friends with 159 saying that they have more than 500 friends. Use the "plus-four" method to find a 90% confidence interval for the true proportion of teens that would report having more than 500 Facebook friends based on this larger sample. Compare the results to those in Example.

Answer

Solution A

Using "plus-four," we have $x = 159 + 2 = 161$ and $n = 588 + 4 = 592$.

$$p' = 161592 \approx 0.272$$

$$q' = 1 - p' = 1 - 0.272 = 0.728$$

Since *CL* = 0.90, we know $\alpha = 1 - 0.90 = 0.10$ and $\frac{\alpha}{2} = 0.05$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.272)(0.728)}{592}}\right) \approx 0.030$$

$$p' - EPB = 0.272 - 0.030 = 0.242$$

$$p' + EPB = 0.272 + 0.030 = 0.302$$

We are 90% confident that between 24.2% and 30.2% of all teens would report having more than 500 friends on Facebook.

Solution B

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint. Press ENTER.
- Arrow down to x and enter 161.
- Arrow down to n and enter 592.
- Arrow down to C-Level and enter 0.90.
- Arrow down to Calculate and press ENTER.
- The confidence interval is (0.242, 0.302).

Conclusion: The confidence interval for the larger sample is narrower than the interval from Example. Larger samples will always yield more precise confidence intervals than smaller samples. The "plus four" method has a greater impact on the smaller sample. It shifts the point estimate from 0.26 (13/50) to 0.278 (15/54). It has a smaller impact on the *EPB*, changing it from 0.102 to 0.100. In the larger sample, the point estimate undergoes a smaller shift: from 0.270 (159/588) to 0.272 (161/592). It is easy to see that the plus-four method has the greatest impact on smaller samples.

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population proportion is

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right)$$

Solving for n gives you an equation for the sample size.

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 (p'q')}{EBP^2}$$

Example 9.4.5

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Answer

From the problem, we know that $\mathbf{EBP = 0.03}$ (3%=0.03) and $z_{\frac{\alpha}{2}}$ $z_{0.05} = 1.645$ because the confidence level is 90%.

However, in order to find n, we need to know the estimated (sample) proportion p'. Remember that $q' = 1 - p'$. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because $p'q' = (0.5)(0.5) = 0.25$ results in the largest possible product. (Try other products: $(0.6)(0.4) = 0.24$; $(0.3)(0.7) = 0.21$; $(0.2)(0.8) = 0.16$ and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$$n = \frac{z^2 p'q'}{EBP^2}$$

gives

$$n = \frac{1.645^2 (0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Exercise 9.4.5

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

Answer

271 customers should be surveyed. Check the Real Estate section in your local

Glossary

Binomial Distribution

a discrete random variable (RV) which arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all

trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(\mathbf{n}, \mathbf{p})$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x = \binom{n}{x})p^x q^{n-x}$.

Error Bound for a Population Proportion (EBP)

the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

9.5: Confidence Interval - Home Costs (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will calculate the 90% confidence interval for the mean cost of a home in the area in which this school is located.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

Check the Real Estate section in your local newspaper. Record the sale prices for 35 randomly selected homes recently listed in the county.

NOTE

Many newspapers list them only one day per week. Also, we will assume that homes come up for sale randomly.

1. Complete the table:

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Describe the Data

1. Compute the following:
    1. $\bar{x}$ = _____
    2. s_x = _____
    3. n = _____
2. In words, define the random variable $\bar{X}$.
3. State the estimated distribution to use. Use both words and symbols.

Find the Confidence Interval

1. Calculate the confidence interval and the error bound.
    1. Confidence Interval: _____
    2. Error Bound: _____
2. How much area is in both tails (combined)? α = _____
3. How much area is in each tail? $\frac{\alpha}{2}$ = _____
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample mean.

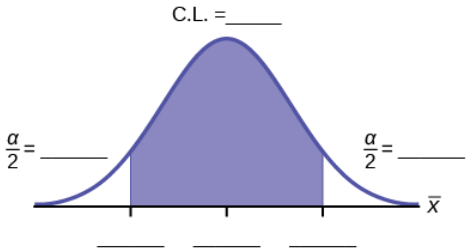


Figure 8.5.1.

5. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data on the first page and count how many of the data values lie within the confidence interval. What percent is this? Is this percent close to 90%? Explain why this percent should or should not be close to 90%.

Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as if you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.

Use the Data to Construct Confidence Intervals

1. Using the given information, construct a confidence interval for each confidence level given.

Confidence level	EBM/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

2. What happens to the EBM as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

9.6: Confidence Interval -Place of Birth (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will calculate the 90% confidence interval the proportion of students in this school who were born in this state.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

1. Survey the students in your class, asking them if they were born in this state. Let $X =$ the number that were born in this state.
    1. $n =$ _____
    2. $x =$ _____
2. In words, define the random variable P'.
3. State the estimated distribution to use.

Find the Confidence Interval and Error Bound

1. Calculate the confidence interval and the error bound.
    1. Confidence Interval: _____
    2. Error Bound: _____
2. How much area is in both tails (combined)? $\alpha =$ _____
3. How much area is in each tail? $\frac{\alpha}{2} =$ _____
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample proportion. <figure >

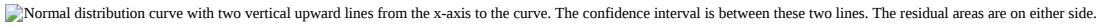


Figure 8.6.1.

Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as though you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.
3. Construct a confidence interval for each confidence level given.

Confidence level	EBP/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

4. What happens to the EBP as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

9.7: Confidence Interval -Women's Heights (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will calculate a 90% confidence interval using the given data.
- The student will determine the relationship between the confidence level and the percentage of constructed intervals that contain the population mean.

Given:

Heights of 100 Women (in Inches)

59.4	71.6	69.3	65.0	62.9	66.5	61.7	55.2
67.5	67.2	63.8	62.9	63.0	63.9	68.7	65.5
61.9	69.6	58.7	63.4	61.8	60.6	69.8	60.0
64.9	66.1	66.8	60.6	65.6	63.8	61.3	59.2
64.1	59.3	64.9	62.4	63.5	60.9	63.3	66.3
61.5	64.3	62.9	60.6	63.8	58.8	64.9	65.7
62.5	70.9	62.9	63.1	62.2	58.7	64.7	66.0
60.5	64.7	65.4	60.2	65.0	64.1	61.1	65.3
64.6	59.2	61.4	62.0	63.5	61.4	65.5	62.3
65.5	64.7	58.8	66.1	64.9	66.9	57.9	69.8
58.5	63.4	69.2	65.9	62.2	60.0	58.1	62.5
62.4	59.1	66.4	61.2	60.4	58.7	66.7	67.5
63.2	56.6	67.7	62.5				

1. Table lists the heights of 100 women. Use a random number generator to select ten data values randomly.
2. Calculate the sample mean and the sample standard deviation. Assume that the population standard deviation is known to be 3.3 inches. With these values, construct a 90% confidence interval for your sample of ten values. Write the confidence interval you obtained in the first space of Table.
3. Now write your confidence interval on the board. As others in the class write their confidence intervals on the board, copy them into Table.

90% Confidence Intervals

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Discussion Questions

1. The actual population mean for the 100 heights given Table is $\mu = 63.4$. Using the class listing of confidence intervals, count how many of them contain the population mean μ; i.e., for how many intervals does the value of μ lie between the endpoints of the confidence interval?
2. Divide this number by the total number of confidence intervals generated by the class to determine the percent of confidence intervals that contains the mean μ. Write this percent here: _____.
3. Is the percent of confidence intervals that contain the population mean μ close to 90%?
4. Suppose we had generated 100 confidence intervals. What do you think would happen to the percent of confidence intervals that contained the population mean?
5. When we construct a 90% confidence interval, we say that we are **90% confident that the true population mean lies within the confidence interval.** Using complete sentences, explain what we mean by this phrase.
6. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data given (the heights of women) and count how many of the data values lie within the confidence interval that you generated based on that data. How many of the 100 data values lie within your confidence interval? What percent is this? Is this percent close to 90%?
7. Explain why it does not make sense to count data values that lie in a confidence interval. Think about the random variable that is being used in the problem.
8. Suppose you obtained the heights of ten women and calculated a confidence interval from this information. Without knowing the population mean μ, would you have any way of knowing **for certain** if your interval actually contained the value of μ ? Explain.

9.E: Confidence Intervals (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

8.1: Introduction

8.2: A Single Population Mean using the Normal Distribution

Q 8.2.1

Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

a. i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. In words, define the random variables X and $\bar{X}$.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean height of male Swedes.

   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the error bound obtained if 1,000 male Swedes are surveyed instead of 48? Why?

S 8.2.1

a. i. 71
   ii. 3
   iii. 48

b. X is the height of a Swedish male, and is the mean height from a sample of 48 Swedish males.

c. Normal. We know the standard deviation for the population, and the sample size is greater than 30.

d. i. CI: (70.151, 71.849)

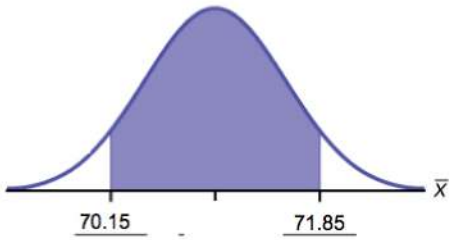


Figure 8.2.1.

   ii. EBM = 0.849

e. The error bound will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

Q 8.2.2

Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.

1. In words, define the random variables X and $\bar{X}$.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 95% confidence interval for the population mean length of engineering conferences.
   1. State the confidence interval.
   2. Sketch the graph.

3. Calculate the error bound.

Q 8.2.3

Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

a. i. $\bar{x} =$ _____
   ii. $\sigma =$ _____
   iii. $n =$ _____
b. In words, define the random variables X and $\bar{X}$.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean time to complete the tax forms.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
f. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
g. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

S 8.2.3

a. i. $\bar{x} = 23.6$
   ii. $\sigma = 7$
   iii. $n = 100$
b. X is the time needed to complete an individual tax form. $\bar{X}$ is the mean time to complete tax forms from a sample of 100 customers.
c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.
d. i. (22.228, 24.972)
   ii.

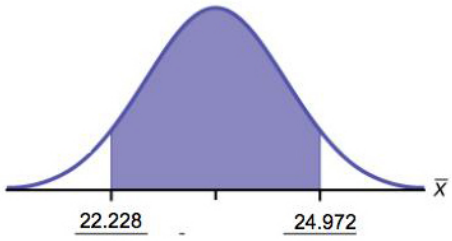


Figure 8.2.7.

   iii. $EBM = 1.372$
e. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.
f. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.
g. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

Q 8.2.4

A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

a. i. $\bar{x} =$_____
   ii. $\sigma =$_____
   iii. $s_x =$_____
b. In words, define the random variable X.
c. In words, define the random variable $\bar{X}$.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 90% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
f. Construct a 98% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
g. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval in part e.
h. In complete sentences, give an interpretation of what the interval in part f means.

Q 8.2.5

A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

a. i. $\bar{x} =$_____
   ii. $\sigma =$_____
   iii. $x =$_____
b. Define the random variables X and $\bar{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90% confidence interval for the population mean number of letters campers send home.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

S 8.2.5

a. i. 7.9
   ii. 2.5
   iii. 20
b. X is the number of letters a single camper will send home. $\bar{X}$ is the mean number of letters sent home from a sample of 20 campers.
c. $N7.9\left(\frac{2.5}{\sqrt{20}}\right)$
d.
   i. CI: (6.98, 8.82)

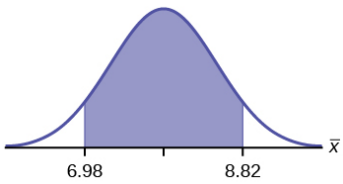


Figure 8..2.2: Copy and Paste Caption here. (Copyright; author via source)

   i. EBM : 0.92

e. The error bound and confidence interval will decrease.

Q 8.2.6

What is meant by the term “90% confident” when constructing a confidence interval for a mean?

1. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
2. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
3. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
4. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

Q 8.2.7

The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. Table shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $100. The standard deviation for this data to the nearest hundred is σ = $909,200.

$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200
$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

a. Find the point estimate for the population mean.
b. Using 95% confidence, calculate the error bound.
c. Create a 95% confidence interval for the mean total individual contributions.
d. Interpret the confidence interval in the context of the problem.

S 8.2.7

a. $\bar{x} = \$568,873$
b. $CL = 0.95 \alpha = 1 - 0.95 = 0.05 z_{\frac{\alpha}{2}} = 1.96$
$EBM = z_{0.025} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{909200}{\sqrt{40}} = \$281,764$
c. $\bar{x}$ − *EBM* = 568,873 − 281,764 = 287,109
$\bar{x}$ + *EBM* = 568,873 + 281,764 = 850,637

Alternate solution:

1. Press `STAT` and arrow over to `TESTS` .
2. Arrow down to `7:ZInterval` .
3. Press `ENTER` .
4. Arrow to Stats and press `ENTER` .
5. Arrow down and enter the following values:
   - $\sigma : 909,200$
   - $\bar{x} : 568,873$
   - $n : 40$
   - $CL : 0.95$

6. Arrow down to Calculate and press ENTER .
7. The confidence interval is ($287,114, $850,632).
8. Notice the small difference between the two solutions–these differences are simply due to rounding error in the hand calculations.

d. We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between $287,109 and $850,637.

Q 8.2.8

The American Community Survey (ACS), part of the United States Census Bureau, conducts a yearly census similar to the one taken every ten years, but with a smaller percentage of participants. The most recent survey estimates with 90% confidence that the mean household income in the U.S. falls between $69,720 and $69,922. Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

Q 8.2.9

The average height of young adult males has a normal distribution with standard deviation of 2.5 inches. You want to estimate the mean height of students at your college or university to within one inch with 93% confidence. How many male students must you measure?

S 8.2.9

Use the formula for EBM, solved for n:

$n = \frac{z^2\sigma^2}{EBM^2}$

From the statement of the problem, you know that σ = 2.5, and you need $EBM = 1$.

$z = z_{0.035} = 1.812$

(This is the value of z for which the area under the density curve to the ***right*** of z is 0.035.)

$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$

You need to measure at least 21 male students to achieve your goal.

8.3: A Single Population Mean using the Student t Distribution

Q 8.3.1

In six packages of “The Flintstones® Real Fruit Snacks” there were five Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice
c. Calculate p'.
d. Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

Q 8.3.2

A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

a. i. $\bar{x} =$ _____
   ii. $s_x =$ _____
   iii. $n =$ _____

iv. $n-1=$ _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.

i. State the confidence interval.

ii. Sketch the graph.

iii. Calculate the error bound.

e. What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

S 8.3.2

a. i. 8629

ii. 6944

iii. 35

iv. 34

b. t_{34}

c. i. $CI: (6244, 11,014)$

ii.

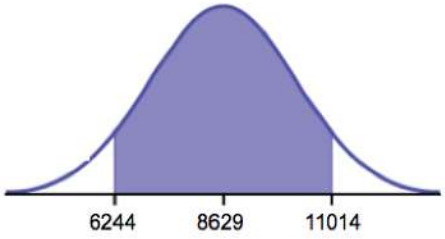


Figure 8.3.4.

iii. $EB = 2385$

d. It will become smaller

Q 8.3.3

Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

a. i. $\bar{x} =$ _____

ii. $s_x =$ _____

iii. $n =$ _____

iv. $n-1=$ _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean time wasted.

i. State the confidence interval.

ii. Sketch the graph.

iii. Calculate the error bound.

e. Explain in a complete sentence what the confidence interval means.

Q 8.3.4

A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4.

a. i. $\bar{x} =$ _____

ii. $s_x =$ _____

iii. $n =$ _____

iv. $n-1 =$ _____

b. Define the random variable X in words.
c. Define the random variable $\bar{X}$ in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 95% confidence interval for the population mean length of time.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
f. What does it mean to be "95% confident" in this problem?

S 8.3.4

a. i. $\bar{x} = 2.51$
   ii. $s_x = 0.318$
   iii. $n = 9$
   iv. $n - 1 = 8$
b. the effective length of time for a tranquilizer
c. the mean effective length of time of tranquilizers from a sample of nine patients
d. We need to use a Student's-t distribution, because we do not know the population standard deviation.
e. i. $CI : (2.27, 2.76)$
   ii. Check student's solution.
   iii. $EBM : 0.25$
f. If we were to sample many groups of nine patients, 95% of the samples would contain the true population mean length of time.

Q 8.3.5

Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

a. i. $\bar{x} =$ _____
   ii. $s_x =$ _____
   iii. $n =$ _____
   iv. $n-1 =$ _____
b. Define the random variable X in words.
c. Define the random variable $\bar{X}$ in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 99% confidence interval for the population mean length of time using training wheels.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
f. Why would the error bound change if the confidence level were lowered to 90%?

Q 8.3.6

The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.

The FEC has reported financial information for 556 Leadership PACs that operating during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 20 Leadership PACs.

$46,500.00	$0	$40,966.50	$105,887.20	$5,175.00
$29,050.00	$19,500.00	$181,557.20	$31,500.00	$149,970.80

$2,555,363.20	$12,025.00	$409,000.00	$60,521.70	$18,000.00
$61,810.20	$76,530.80	$119,459.20	$0	$63,520.00
$6,500.00	$502,578.00	$705,061.10	$708,258.90	$135,810.00
$2,000.00	$2,000.00	$0	$1,287,933.80	$219,148.30

$\bar{x} = \$251,854.23$

$s = \$521,130.41$

Use this sample data to construct a 96% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's t-distribution.

S 8.3.6

$\bar{x} = \$251,854.23$

$s = \$521,130.41$

Note that we are not given the population standard deviation, only the standard deviation of the sample.

There are 30 measures in the sample, so $n = 30$, and $df = 30 - 1 = 29$

$CL = 0.96$, so $\alpha = 1 - CL = 1 - 0.96 = 0.04$

$\frac{\alpha}{2} = 0.02 t_{0.02} = t_{0.02} = 2.150$

$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 2.150 \left(\frac{521,130.41}{\sqrt{30}} \right) - \$204,561.66$

$\bar{x} - EBM = \$251,854.23 - \$204,561.66 = \$47,292.57$

$\bar{x} + EBM = \$251,854.23 + \$204,561.66 = \$456,415.89$

We estimate with 96% confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between $47,292.57 and $456,415.89.

Alternate Solution

Enter the data as a list.

Press `STAT` and arrow over to `TESTS` .

Arrow down to `8:TInterval` .

Press `ENTER` .

Arrow to Data and press `ENTER` .

Arrow down and enter the name of the list where the data is stored.

Enter `Freq` : 1

Enter `C-Level` : 0.96

Arrow down to `Calculate` and press `Enter` .

The 96% confidence interval is ($47,262, $456,447).

The difference between solutions arises from rounding differences.

Q 8.3.7

Forbes magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. The Table shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56

59	74	63	53	50
59	60	60	57	46
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Use this sample data to construct a 90% confidence interval for the mean age of CEO's for these top small firms. Use the Student's t-distribution.

Q 8.3.8

Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

a. i. $\bar{x} =$ _____
   ii. $s_x =$ _____
   iii. $n =$ _____
   iv. $n - 1 =$ _____
b. Define the random variables X and $\bar{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

S 8.3.8

a. i. $\bar{x} =$
   ii. $s_x =$
   iii. $n =$
   iv. $n - 1 =$
b. X is the number of unoccupied seats on a single flight. $\bar{X}$ is the mean number of unoccupied seats from a sample of 225 flights.
c. We will use a Student's t-distribution, because we do not know the population standard deviation.
d. i. $CI : (11.12, 12.08)$
   ii. Check student's solution.
   iii. EBM: 0.48

Q 8.3.9

In a recent sample of 84 used car sales costs, the sample mean was $6,425 with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

a. Which distribution should you use for this problem? Explain your choice.
b. Define the random variable $\bar{X}$ in words.
c. Construct a 95% confidence interval for the population mean cost of a used car.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
d. Explain what a "95% confidence interval" means for this study.

Q 8.3.10

Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

a. Construct a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
b. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
c. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
d. Calculate the mean.
e. Is the mean within the interval you calculated in part a? Did you expect it to be? Why or why not?

S 8.3.10

a. i. CI: (7.64 , 9.36)

ii.

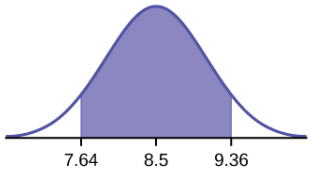


Figure 8.3.5.

iii. EBM : 0.86

b. The sample should have been increased.
c. Answers will vary.
d. Answers will vary.
e. Answers will vary.

Q 8.3.11

A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; $1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

a. i. $\bar{x} =$ _____
   ii. $s_x =$ _____
   iii. $n =$ _____
   iv. $n - 1 =$ _____
b. Define the random variables X and $\bar{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean worth of coupons.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

Use the following information to answer the next two exercises: A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

Q 8.3.12

Find the 95% Confidence Interval for the true population mean for the amount of soda served.

a. (12.42, 14.18)
b. (12.32, 14.29)
c. (12.50, 14.10)
d. Impossible to determine

S 8.3.12

b

Q 8.3.13

What is the error bound?

a. 0.87
b. 1.98
c. 0.99
d. 1.74

8.4: A Population Proportion

Q 8.4.1

Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
b. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

S 8.4.1

a. 1,068
b. The sample size would need to be increased since the critical value increases as the confidence level increases.

Q 8.4.2

Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

a. i. $x =$ _____
   ii. $n =$ _____
   iii. $p' =$ _____
b. Define the random variables X and P', in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population proportion who claim they always buckle up.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

Q 8.4.3

According to a recent survey of 1,200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.
   i. State the confidence interval.

ii. Sketch the graph.
iii. Calculate the error bound.

S 8.4.3

a. $X =$ the number of people who feel that the president is doing an acceptable job;

$P' =$ the proportion of people in a sample who feel that the president is doing an acceptable job.

b. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$

c. i. $CI: (0.59, 0.63)$
ii. Check student's solution
iii. $EBM: 0.02$

Q 8.4.4

An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1,709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites. In this survey, 86% of blacks said that they would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person.

a. We are interested in finding the 95% confidence interval for the percent of all black adults who would welcome a white person into their families. Define the random variables X and P', in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

Q 8.4.5

Refer to the information in Exercise.

a. Construct three 95% confidence intervals.
   i. percent of all Asians who would welcome a white person into their families.
   ii. percent of all Asians who would welcome a Latino into their families.
   iii. percent of all Asians who would welcome a black person into their families.
b. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
c. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
d. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

S 8.4.5

a. i. (0.72, 0.82)
   ii. (0.65, 0.76)
   iii. (0.60, 0.72)
b. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.
c. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.
d. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.

Q 8.4.6

Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
d. Explain what a "97% confidence interval" means for this study.

Q 8.4.7

A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was "What is the main problem facing the country?" Twenty percent answered "crime." We are interested in the population proportion of adult Americans who feel that crime is the main problem.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
d. Suppose we want to lower the sampling error. What is one way to accomplish that?
e. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ± 3. In one to three complete sentences, explain what the ±3% represents.

S 8.4.7

a. $X =$ the number of adult Americans who feel that crime is the main problem; $P' =$ the proportion of adult Americans who feel that crime is the main problem
b. Since we are estimating a proportion, given $P' = 0.2$ and $n = 1000$, the distribution we should use is $N\left(0.61, \sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.
c. i. $CI : (0.18, 0.22)$
   ii. Check student's solution.
   iii. $EBM : 0.02$
d. One way to lower the sampling error is to increase the sample size.
e. The stated "± 3" represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3%. Thus, they estimate the percentage of adult Americans who feel that crime is the main problem to be between 18% and 22%.

Q 8.4.8

Refer to Exercise. Another question in the poll was "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
   i. State the confidence interval.
   ii. Sketch the graph.

iii. Calculate the error bound.

d. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ± 3. In one to three complete sentences, explain what the ±3% represents.

Use the following information to answer the next three exercises: According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that "education and our schools" is one of the top issues facing California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

Q 8.4.9

A point estimate for the true population proportion is:

a. 0.90
b. 1.27
c. 0.79
d. 400

S 8.4.9

c

Q 8.4.10

A 90% confidence interval for the population proportion is _____.

a. (0.761, 0.820)
b. (0.125, 0.188)
c. (0.755, 0.826)
d. (0.130, 0.183)

Q 8.4.11

The error bound is approximately _____.

a. 1.581
b. 0.791
c. 0.059
d. 0.030

S 8.4.11

d

Use the following information to answer the next two exercises: Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness, and 338 did not.

Q 8.4.12

Find the confidence interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

a. (0.2975, 0.3796)
b. (0.6270, 0.6959)
c. (0.3041, 0.3730)
d. (0.6204, 0.7025)

Q 8.4.13

The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is _____.

a. 0.6614
b. 0.3386
c. 173
d. 338

S 8.4.13

a

Q 8.4.14

On May 23, 2013, Gallup reported that of the 1,005 people surveyed, 76% of U.S. workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95% with a ± 3 margin of error.

a. Determine the estimated proportion from the sample.
b. Determine the sample size.
c. Identify CL and α.
d. Calculate the error bound based on the information provided.
e. Compare the error bound in part d to the margin of error reported by Gallup. Explain any differences between the values.
f. Create a confidence interval for the results of this study.
g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

Q 8.4.15

A national survey of 1,000 adults was conducted on May 13, 2013 by Rasmussen Reports. It concluded with 95% confidence that 49% to 55% of Americans believe that big-time college sports programs corrupt the process of higher education.

a. Find the point estimate and the error bound for this confidence interval.
b. Can we (with 95% confidence) conclude that more than half of all American adults believe this?
c. Use the point estimate from part a and $n = 1,000$ to calculate a 75% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.
d. Can we (with 75% confidence) conclude that at least half of all American adults believe this?

S 8.4.15

a. $p' = \frac{(0.55+0.49)}{2} = 0.52; EBP = 0.55 - 0.52 = 0.03$
b. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.
c. $CL = 0.75$, so $\alpha = 1 - 0.75 = 0.25$ and $\frac{\alpha}{2} = 0.125 z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is $1 - 0.125 = 0.875$)
$EBP = (1.150)\sqrt{\frac{0.52(0.48)}{1,000}} \approx 0.018$
$(p' - EBP, p' + EBP) = (0.52 - 0.018, 0.52 + 0.018) = (0.502, 0.538)$

Alternate Solution

STAT TESTS A: 1-PropZinterval with $x = (0.52)(1,000), n = 1,000, CL = 0.75$

Answer is (0.502, 0.538)

d. Yes – this interval does not fall less than 0.50 so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75% confidence.

Q 8.4.16

Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

a. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.
b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.

c. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

Q 8.4.17

You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

S 8.4.17

$CL = 0.95 \alpha = 1 - 0.95 = 0.05 \frac{\alpha}{2} = 0.025 z_{\frac{\alpha}{2}} = 1.96$. Use $p' = q' = 0.5$.

$$n = \frac{z^2_{\frac{\alpha}{2}} p' q'}{EPB^2} = \frac{1.96^2 (0.5)(0.5)}{0.05^2} = 384.16$$

You need to interview at least 385 students to estimate the proportion to within 5% at 95% confidence.

Q 8.4.18

In a recent Zogby International Poll, nine of 48 respondents rated the likelihood of a terrorist attack in their community as “likely” or “very likely.” Use the “plus four” method to create a 97% confidence interval for the proportion of American adults who believe that a terrorist attack in their community is likely or very likely. Explain what this confidence interval means in the context of the problem.

8.5: Confidence Interval (Home Costs)

8.6: Confidence Interval (Place of Birth)

8.7: Confidence Interval (Women's Heights)

9.S: Confidence Intervals (Summary)

Review

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form:

$$(\text{lower bound, upper bound}) = (\text{point estimate} - EBM, \text{point estimate} + EBM)$$

The calculation of EBM depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding EBM increases as well. As the sample size increases, the EBM decreases. By the central limit theorem,

$$EBM = z\frac{\sigma}{\sqrt{n}}$$

Given a confidence interval, you can work backwards to find the error bound (EBM) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the EBM formula for n to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{z^2\sigma^2}{EBM^2}$$

Formula Review

$\bar{X} - N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by

$$(\text{lower bound, upper bound}) = (\text{point estimate} - EBM, \text{point estimate} + EBM) \tag{9.S.1}$$

$$= \bar{x} - EBM, \bar{x} + EBM \tag{9.S.2}$$

$$= \left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right) \tag{9.S.3}$$

$EBM = z\frac{\sigma}{\sqrt{n}} =$ the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

$CL =$ confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1 - CL =$ the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the z-score with the property that the area to the right of the z-score is $\frac{\alpha}{2}$ this is the z-score used in the calculation of "EBM where $\alpha = 1 - CL$".

$n = \frac{z^2\sigma^2}{EBM^2}$ the formula used to determine the sample size (n) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

$$(\text{lower value, upper value}) = (\text{point estimate} - \text{error bound, point estimate} + \text{error bound}) \tag{9.S.4}$$

To find the error bound when you know the confidence interval

$$\text{error bound} = \text{upper value} - \text{point estimate} \tag{9.S.5}$$

OR

$$\text{error bound} = \frac{\text{upper value - lower value}}{2} \tag{9.S.6}$$

Single Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means, Population Standard Deviation is Known $EBM = z\frac{\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

? Exercise 8.2.8

Identify the following:

a. $\bar{x} =$ _____
b. $\sigma =$ _____
c. $n =$ _____

Answer

a. 244
b. 15
c. 50

? Exercise 8.2.9

In words, define the random variables X and $\bar{X}$.

? Exercise 8.2.10

Which distribution should you use for this problem?

Answer

$N\left(244, \frac{15}{\sqrt{50}}\right)$

? Exercise 8.2.11

Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

? Exercise 8.2.12

What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Answer

As the sample size increases, there will be less variability in the mean, so the interval size decreases.

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

Exercise 8.2.13

Identify the following:

a. $\bar{x} =$ _____
b. $\sigma =$ _____
c. $n =$ _____

Exercise 8.2.14

In words, define the random variables X and $\bar{X}$.

Answer

X is the time in minutes it takes to complete the U.S. Census short form. $\bar{X}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

Exercise 8.2.15

Which distribution should you use for this problem?

Exercise 8.2.16

Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$CI : (7.9441, 8.4559)$

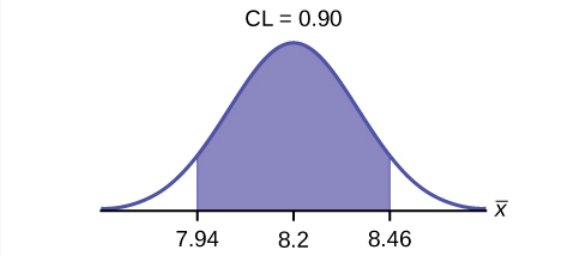


Figure 9.S.3.

$EBM = 0.26$

Exercise 8.2.17

If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

Exercise 8.2.18

If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

Answer

The level of confidence would decrease because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

Exercise 8.2.19

Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

Exercise 8.2.20

Identify the following:

a. $\bar{x} =$ _____
b. $\sigma =$ _____
c. $n =$ _____

Answer

a. $\bar{x} = 2.2$
b. $\sigma = 0.2$
c. $n = 20$

Exercise 8.2.21

In words, define the random variable X.

Exercise 8.2.22

In words, define the random variable $\bar{X}$.

Answer

$\bar{X}$ is the mean weight of a sample of 20 heads of lettuce.

Exercise 8.2.23

Which distribution should you use for this problem?

Exercise 8.2.24

Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$EBM = 0.07$

$CI : (2.1264, 2.2736)$

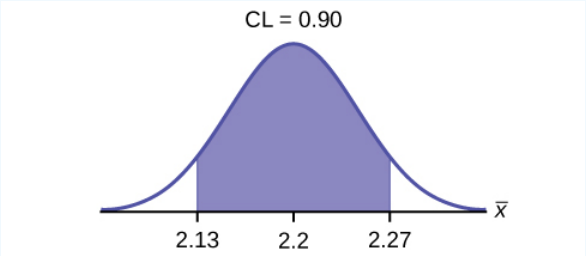


Figure 9.S.4.

Exercise 8.2.25

Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

Exercise 8.2.26

In complete sentences, explain why the confidence interval in Exercise is larger than in Exercise.

Answer

The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

Exercise 8.2.27

In complete sentences, give an interpretation of what the interval in Exercise means.

Exercise 8.2.28

What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

Answer

The confidence level would increase.

Exercise 8.2.29

What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

Exercise 8.2.30

$\bar{x} =$ _____

Answer

30.4

Exercise 8.2.31

$n =$ _____

Exercise 8.2.32

_____ $= 15$

Answer

σ

Exercise 8.2.33

In words, define the random variable $\bar{X}$.

Exercise 8.2.34

What is $\bar{x}$ estimating?

Answer

μ

Exercise 8.2.35

Is σ_x known?

Exercise 8.2.36

As a result of your answer to Exercise, state the exact distribution to use when calculating the confidence interval.

Answer

normal

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

Exercise 8.2.37

How much area is in both tails (combined)? $\alpha =$ _____

Exercise 8.2.38

How much area is in each tail? $\frac{\alpha}{2} =$ _____

Answer

0.025

Exercise 8.2.39

Identify the following specifications:

a. lower limit
b. upper limit
c. error bound

Exercise 8.2.40

The 95% confidence interval is:_____.

Answer

(24.52,36.28)

Exercise 8.2.41

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

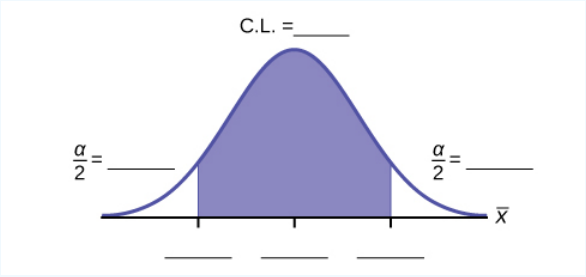


Figure 9.S.5.

Exercise 8.2.42

In one complete sentence, explain what the interval means.

Answer

We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

Exercise 8.2.43

Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

Exercise 8.2.44

Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

Answer

The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

Review

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, s, as an estimate of σ. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the Student's t-distribution is much better. Define a t-score using the following formula:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{9.S.7}$$

The t-score follows the Student's t-distribution with $n-1$ degrees of freedom. The confidence interval under this distribution is calculated with $EBM = \left(t_{\frac{\alpha}{2}}\right)\frac{s}{\sqrt{n}}$ where $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

Formula Review

$s =$ the standard deviation of sample values.

$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ is the formula for the t-score which measures how far away a measure is from the population mean in the Student's t-distribution

$df = n - 1$; the degrees of freedom for a Student's t-distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T, has a Student's t-distribution with df degrees of freedom

$EBM = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} =$ the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the t-score in the Student's t-distribution with area to the right equal to $\frac{\alpha}{2}$

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by (lower bound, upper bound)

$$= (\text{point estimate} - EBM, \text{point estimate} + EBM) \tag{9.S.8}$$

$$= \left(\bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}}\right) \tag{9.S.9}$$

Use the following information to answer the next five exercises. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

? Exercise 8.3.3

Identify the following:

a. $\bar{x} =$_____
b. $s_x =$_____
c. $n =$_____
d. $n - 1 =$_____

? Exercise 8.3.4

Define the random variables X and $\bar{X}$ in words.

Answer

X is the number of hours a patient waits in the emergency room before being called back to be examined. $\bar{X}$ is the mean wait time of 70 patients in the emergency room.

? Exercise 8.3.5

Which distribution should you use for this problem?

? Exercise 8.3.6

Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$CI : (1.3808, 1.6192)$

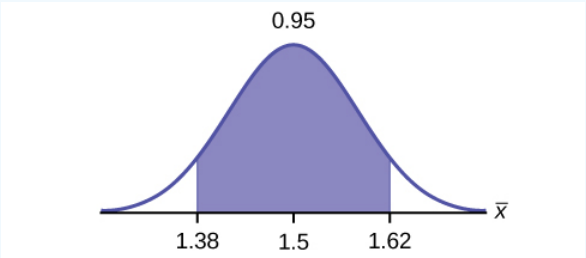


Figure 9.S.1.

$EBM = 0.12$

Exercise 8.3.7

Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

Exercise 8.3.8

a. $\bar{x} =$_____
b. $s_x =$_____
c. $n =$_____
d. $n - 1 =$_____

Answer

a. $\bar{x} = 151$
b. $s_x = 32$
c. $n = 108$
d. $n - 1 = 107$

Exercise 8.3.9

Define the random variable X in words.

Exercise 8.3.10

Define the random variable $\bar{X}$ in words.

Answer

$\bar{X}$ is the mean number of hours spent watching television per month from a sample of 108 Americans.

Exercise 8.3.11

Which distribution should you use for this problem?

Exercise 8.3.12

Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.

Answer

$CI : (142.92, 159.08)$

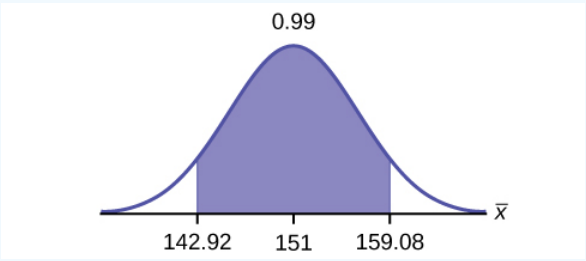


Figure 9.S.2.

$EBM = 8.08$

Exercise 8.3.13

Why would the error bound change if the confidence level were lowered to 95%?

Use the following information to answer the next 13 exercises: The data in Table are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let $X =$ the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

Exercise 8.3.14

a. $\bar{x} =$_____
b. $s_x =$_____
c. $n =$_____

Answer

a. 3.26
b. 1.02
c. 39

Exercise 8.3.15

Define the random variable $\bar{X}$ in words.

Exercise 8.3.16

What is $\bar{x}$ estimating?

Answer

μ

Exercise 8.3.17

Is σ_x known?

Exercise 8.3.18

As a result of your answer to Exercise, state the exact distribution to use when calculating the confidence interval.

Answer

t_{38}

Construct a 95% confidence interval for the true mean number of colors on national flags.

Exercise 8.3.19

How much area is in both tails (combined)?

Exercise 8.3.20

How much area is in each tail?

Answer

0.025

Exercise 8.3.21

Calculate the following:

a. lower limit
b. upper limit
c. error bound

Exercise 8.3.22

The 95% confidence interval is_____.

Answer

(2.93, 3.59)

Exercise 8.3.23

Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.

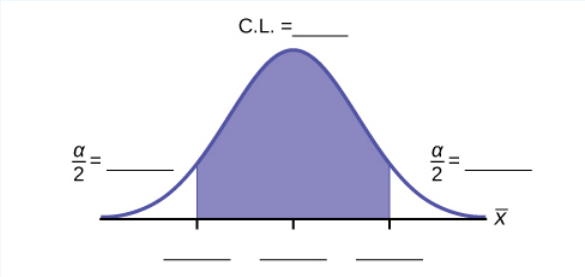


Figure 9.S.3.

Exercise 8.3.24

In one complete sentence, explain what the interval means.

Answer

We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

Exercise 8.3.25

Using the same $\bar{x}$, s_x, and level of confidence, suppose that n were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

Answer

The error bound would become $EBM = 0.245$. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

? Exercise 8.3.26

Using the same $\bar{x}$, s_x, and $n = 39$, how would the error bound change if the confidence level were reduced to 90%? Why?

References

1. Jensen, Tom. "Democrats, Republicans Divided on Opinion of Music Icons." Public Policy Polling. Available online at www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).
2. Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. "Teens, Social Media, and Privacy." PewInternet, 2013. Available online at www.pewinternet.org/Reports/2...d-Privacy.aspx (accessed July 2, 2013).
3. Prince Survey Research Associates International. "2013 Teen and Privacy Management Survey." Pew Research Center: Internet and American Life Project. Available online at www.pewinternet.org/~/media//...al%20Media.pdf (accessed July 2, 2013).
4. Saad, Lydia. "Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity." Gallup® Economy, 2013. Available online at http://www.gallup.com/poll/162758/th...ement-age.aspx (accessed July 2, 2013).
5. The Field Poll. Available online at field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).
6. Zogby. "New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure 'Investment' for National Security." Zogby Analytics, 2013. Available online at http://www.zogbyanalytics.com/news/2...analytics-poll (accessed July 2, 2013).
7. "52% Say Big-Time College Athletics Corrupt Education Process." Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/publ...cation_process (accessed July 2, 2013).

Review

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, $\frac{x}{n}$, where x represents the number of successes and n represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

$$\text{(lower bound, upper bound)} = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}, p' + z\sqrt{\frac{p'q'}{n}}\right)$$

The "plus four" method for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four additional trials in the study; two are successes and two are failures. Calculate $p' = \frac{x+2}{n_4}$, and proceed to find the confidence interval. When sample sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

Formula Review

$p' = \frac{x}{n}$ where x represents the number of successes and n represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$$q' = 1 - p' \tag{9.S.10}$$

$$p' - N\left(p, \sqrt{\frac{pq}{n}}\right) \tag{9.S.11}$$

The variable p' has a binomial distribution that can be approximated with the normal distribution shown here.

$EBP =$ the error bound for a proportion $= z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}}$

Confidence interval for a proportion:

(lower bound, upper bound) $= (p'-EBP, p'+EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}\right), p' + z\sqrt{\frac{p'q'}{n}}$

$n = \frac{z_{\frac{\alpha}{2}}^2 p'q'}{EBP^2}$ provides the number of participants needed to estimate the population proportion with confidence $1-\alpha$ and margin of error EBP.

Use the normal distribution for a single population proportion $p' = \frac{x}{n}$

$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} p' + q' = 1$

The confidence interval has the format $(p'-EBP, p'+EBP)$.

- $\bar{x}$ is a point estimate for μ
- p' is a point estimate for ρ
- s is a point estimate for σ

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

? Exercise 8.4.6

When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?

? Exercise 8.4.7

If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

Answer

It would decrease, because the z-score would decrease, which reducing the numerator and lowering the number.

Use the following information to answer the next five exercises: Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

? Exercise 8.4.8

Identify the following:

a. $x =$ _____
b. $n =$ _____
c. $p' =$ _____

? Exercise 8.4.9

Define the random variables X and P' in words.

Answer

X is the number of "successes" where the woman makes the majority of the purchasing decisions for the household. P' is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

? Exercise 8.4.10

Which distribution should you use for this problem?

Exercise 8.4.11

Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

CI : $(0.5321, 0.6679)$

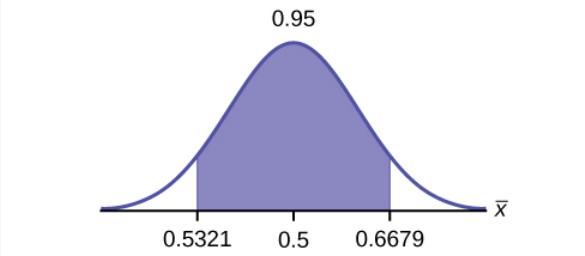


Figure 9.S.1.

EBM : 0.0679

Exercise 8.4.12

List two difficulties the company might have in obtaining random results, if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

Exercise 8.4.13

We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and P' in words.

Answer

X is the number of "successes" where an executive prefers a truck. P' is the percentage of executives sampled who prefer a truck.

Exercise 8.4.14

Which distribution should you use for this problem?

Exercise 8.4.15

Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

CI : $(0.19432, 0.33068)$

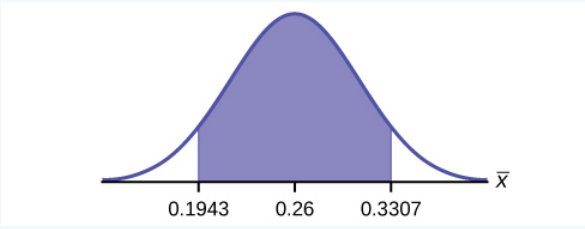


Figure 9.S.2.

EBM : 0.0707

Exercise 8.4.16

Suppose we want to lower the sampling error. What is one way to accomplish that?

Exercise 8.4.17

The sampling error given in the survey is ± 2. Explain what the ± 2 means.

Answer

The sampling error means that the true mean can be 2% above or below the sample mean.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

Exercise 8.4.18

Define the random variable X in words.

Exercise 8.4.19

Define the random variable P' in words.

Answer

P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

Exercise 8.4.20

Which distribution should you use for this problem?

Exercise 8.4.21

Construct a 90% confidence interval, and state the confidence interval and the error bound.

Answer

CI : $(0.62735, 0.67265)$

EBM : 0.02265

Exercise 8.4.22

What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

Exercise 8.4.23

What is being counted?

Answer

The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

Exercise 8.4.24

In words, define the random variable X.

Exercise 8.4.25

Calculate the following:

a. $x =$ _____
b. $n =$ _____
c. $p' =$ _____

Answer

a. $x = 64$
b. $n = 80$
c. $p' = 0.8$

Exercise 8.4.26

State the estimated distribution of X. $X \sim$_____

Exercise 8.4.27

Define a new random variable P'. What is p' estimating?

Answer

p

Exercise 8.4.28

In words, define the random variable P'.

Exercise 8.4.29

State the estimated distribution of P'. Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

Answer

$P - N\left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}}\right)$. $(0.72171, 0.87829)$

Exercise 8.4.30

How much area is in both tails (combined)?

Exercise 8.4.31

How much area is in each tail?

Answer

0.04

Exercise 8.4.32

Calculate the following:

a. lower limit
b. upper limit
c. error bound

Exercise 8.4.33

The 92% confidence interval is _____.

Answer

(0.72; 0.88)

Exercise 8.4.34

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

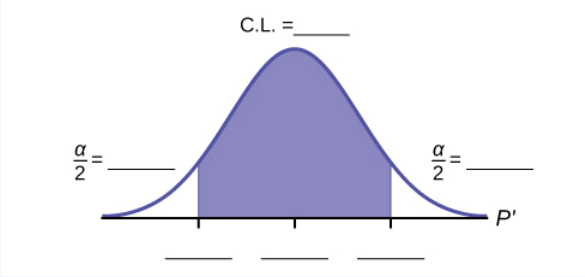


Figure 9.S.3.

Exercise 8.4.35

In one complete sentence, explain what the interval means.

Answer

With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

Exercise 8.4.36

Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

Exercise 8.4.37

Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?

Answer

The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

? Exercise 8.4.38

If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

CHAPTER OVERVIEW

10: Hypothesis Testing with One Sample

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

10.1: Prelude to Hypothesis Testing
10.2: Null and Alternative Hypotheses
10.3: Outcomes and the Type I and Type II Errors
10.4: Distribution Needed for Hypothesis Testing
10.5: Rare Events, the Sample, Decision and Conclusion
10.6: Additional Information and Full Hypothesis Test Examples
10.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet)
10.E: Hypothesis Testing with One Sample (Exercises)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

10.1: Prelude to Hypothesis Testing

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

Figure 10.1.1: You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

A statistician will make a decision about these claims. This process is called "hypothesis testing." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analysis of the data, to reject the null hypothesis. In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

- Set up two contradictory hypotheses.
- Collect sample data (in homework problems, the data or summary statistics will be given to you).
- Determine the correct distribution to perform the hypothesis test.
- Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
- Make a decision and write a meaningful conclusion.

To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See Appendix E.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Hypothesis Testing

Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

10.2: Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0: **The null hypothesis:** It is a statement of no difference between the variables—they are not related. This can often be considered the status quo and as a result if you cannot accept the null it requires some action.

H_a: **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0. This is usually what the researcher is trying to prove.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are "reject H_0" if the sample information favors the alternative hypothesis or "do not reject H_0" or "decline to reject H_0" if the sample information is insufficient to reject the null hypothesis.

Table 10.2.1: Mathematical Symbols Used in H_0 and H_a:

H_0	H_a
equal (=)	not equal $(\neq)$ **or** greater than (>) **or** less than (<)
greater than or equal to $(\geq)$	less than (<)
less than or equal to $(\geq)$	more than (>)

H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use = in the null hypothesis, even with > or < as the symbol in the alternative hypothesis. This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.

Example 10.2.1

- H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$
- H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > 30$

Exercise 10.2.1

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Answer

- H_0: The drug reduces cholesterol by 25%. $p = 0.25$
- H_a: The drug does not reduce cholesterol by 25%. $p \neq 0.25$

Example 10.2.2

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

- $H_0 : \mu = 2.0$
- $H_a : \mu \neq 2.0$

Exercise 10.2.2

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol $(=, \neq, \geq, <, \leq, >)$ for the null and alternative hypotheses.

- H_0 : μ_66
- H_a : μ_66

Answer

- $H_0 : \mu = 66$
- $H_a : \mu \neq 66$

Example 10.2.3

We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

- $H_0 : \mu \geq 5$
- $H_a : \mu < 5$

Exercise 10.2.3

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol ($=, \neq, \geq, <, \leq, >$) for the null and alternative hypotheses.

a. H_0 : μ_45
b. H_a : μ_45

Answer

a. $H_0 : \mu \geq 45$
b. $H_a : \mu < 45$

Example 10.2.4

In an issue of *U. S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

- $H_0 : p \leq 0.066$
- $H_a : p > 0.066$

Exercise 10.2.4

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol $(=, \neq, \geq, <, \leq, >)$ for the null and alternative hypotheses.

a. H_0 : $p\_0.40$
b. H_a : $p\_0.40$

Answer

a. $H_0 : p = 0.40$
b. $H_a : p > 0.40$

COLLABORATIVE EXERCISE

Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality $(=, \leq \text{ or } \geq)$
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using less than, greater than, or not equals symbols, i.e., $(\neq, >, \text{or } <)$.
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

Formula Review

H_0 and H_a are contradictory.

If H_a has:	equal $(=)$	greater than or equal to $(\geq)$	less than or equal to $(\leq)$
then H_a has:	not equal $(\neq)$ **or** greater than $(>)$ **or** less than $(<)$	less than $(<)$	greater than $(>)$

- If $\alpha \leq p$-value, then do not reject H_0.
- If$\alpha > p$-value, then reject H_0.

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.References

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

Glossary

Hypothesis

a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

10.3: Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	H_0 is Actually True	H_0 is Actually False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

The four possible outcomes in the table are:

1. The decision is **not to reject** H_0 when H_0 **is true (correct decision).**
2. The decision is to **reject** H_0 when H_0 **is true** (incorrect decision known as aType I error).
3. The decision is **not to reject** H_0 when, in fact, H_0 **is false** (incorrect decision known as a Type II error).
4. The decision is to **reject** H_0 when H_0 **is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

- $\alpha =$ probability of a Type I error $= P(\text{Type I error}) =$ probability of rejecting the null hypothesis when the null hypothesis is true.
- $\beta =$ probability of a Type II error $= P(\text{Type II error}) =$ probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The *Power of the Test* is $1 - \beta$. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test. The following are examples of Type I and Type II errors.

Example 10.3.1: Type I vs. Type II errors

Suppose the null hypothesis, H_0, is: Frank's rock climbing equipment is safe.

- **Type I error**: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe.
- **Type II error**: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

$\alpha =$ **probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe.

$\beta =$ **probability** that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Exercise 10.3.1

Suppose the null hypothesis, H_0, is: the blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Answer

- **Type I error**: The researcher thinks the blood cultures do contain traces of pathogen X, when in fact, they do not.
- **Type II error**: The researcher thinks the blood cultures do not contain traces of pathogen X, when in fact, they do.

Example 10.3.2

Suppose the null hypothesis, H_0, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

- **Type I error**: The emergency crew thinks that the victim is dead when, in fact, the victim is alive.
- **Type II error**: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

$\alpha =$ **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive $= P(\text{Type I error})$.

$\beta =$ **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead $= P(\text{Type II error})$.

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Exercise 10.3.2

Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Answer

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick, so he will not get treatment.

Example 10.3.3

It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0, is: It's a Boy Genetic Labs has no effect on gender outcome.

- **Type I error**: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.
- **Type II error**: This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.

The error of greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Exercise 10.3.3

"Red tide" is a bloom of poison-producing algae–a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Answer

In this scenario, an appropriate null hypothesis would be H_0: the mean level of toxins is at most 800μg $H_0 : \mu_0 \leq 800\mu$g.

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most 800μg The DMF continues the harvesting ban.

Type II error: The DMF believes that toxin levels are within acceptable levels (are at least 800 μg) when, in fact, toxin levels are still too high (more than 800μg). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

Example 10.3.4

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is the more serious?

- **Type I**: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.
- **Type II**: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

? Exercise 10.3.4

Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%. Identify the Type I and Type II errors from these four statements.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%
b. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

Answer

Type I error: c

Type II error: b

Summary

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected. The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1 - \beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

Formula Review

- $\alpha =$ probability of a Type I error $= P(\text{Type I error}) =$ probability of rejecting the null hypothesis when the null hypothesis is true.
- $\beta =$ probability of a Type II error $= P(\text{Type II error}) =$ probability of not rejecting the null hypothesis when the null hypothesis is false.

Glossary

Type 1 Error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type 2 Error

The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

10.4: Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a normal distribution or a Student's t-distribution. (Remember, use a Student's t-distribution when the population standard deviation is unknown and the distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a single population mean, the distribution for the test is for *means*:

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \tag{10.4.1}$$

or

$$t_{df} \tag{10.4.2}$$

The population parameter is μ. The estimated value (point estimate) for μ is $\bar{x}$, the sample mean.

If you are testing a single population proportion, the distribution for the test is for proportions or percentages:

$$P' \sim N\left(p, \sqrt{\frac{p-q}{n}}\right) \tag{10.4.3}$$

The population parameter is p. The estimated value (point estimate) for p is p'. $p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

Assumptions

When you perform a **hypothesis test of a single population mean** μ using a Student's t-distribution (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a simple random sample that comes from a population that is approximately normally distributed. You use the sample standard deviation to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean** μ using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion** p, you take a simple random sample from the population. You must meet the conditions for a binomial distribution which are: there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five $(np > 5$ and $nq > 5)$. Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1 - p$.

Summary

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of successes and the mean number of failures satisfy the conditions: $np > 5$ and $nq > 5$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Formula Review

If there is no given preconceived α, then use $\alpha = 0.05$.

Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student's t-test**.
- Single population proportion: **Normal test**.
- For a **single population mean**, we may use a normal distribution with the following mean and standard deviation. Means: $\mu = \mu_{\bar{x}}$ and $sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$
- A **single population proportion**, we may use a normal distribution with the following mean and standard deviation. Proportions: $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$.

Glossary

Binomial Distribution

a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p) \mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Student's t-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t-distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.

10.5: Rare Events, the Sample, Decision and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the null hypothesis). Then you gather sample data randomly. If the sample has properties that would be very *unlikely* to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an assumption—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the p-value. The p-value is the probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.

A large p-value calculated from the data indicates that we should not reject the null hypothesis. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example 10.5.1

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm. and the distribution of heights is normal.

- The null hypothesis could be $H_0 : \mu \leq 15$
- The alternate hypothesis is $H_a : \mu > 15$

The words **"is more than"** translates as a ">" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ **is known** ($\sigma = 0.5cm.$), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16.$$

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.

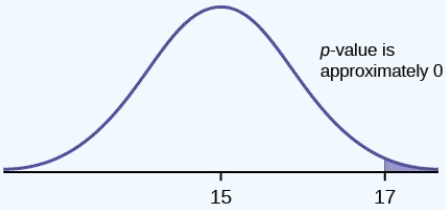


Figure 10.5.1

p-value $= P(\bar{x} > 17)$ which is approximately zero.

A p-value of approximately zero tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Exercise 10.5.1

A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

- $H_0 : \mu \leq 12$
- $H_a : \mu > 12$

The p-value is 0.0013

Draw a graph that shows the p-value.

Answer

p-value $= 0.0013$

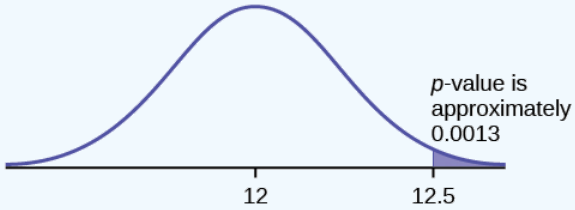


Figure 10.5.2

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the null hypothesis is to compare the p-value and a preset or preconceived α (also called a "**significance level**"). A preset α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a decision to reject or not reject H_0, do as follows:

- If $\alpha > p$-value, reject H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the alternative hypothesis, H_a, may be correct.
- If $\alpha \leq p$-value, do not reject H_0. The results of the sample data are not significant.There is not sufficient evidence to conclude that the alternative hypothesis,H_a, may be correct.

When you "do not reject H_0", it does not mean that you should believe that H_0 is true. It simply means that the sample data have failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_0.

Conclusion: After you make your decision, write a thoughtful conclusion about the hypotheses in terms of the given problem.

Example 10.5.2

When using the p-value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

- If the p-value is low, the null must go.
- If the p-value is high, the null must fly.

This memory aid relates a p-value less than the established alpha (the p is low) as rejecting the null hypothesis and, likewise, relates a p-value higher than the established alpha (the p is high) as not rejecting the null hypothesis.

Fill in the blanks.

Reject the null hypothesis when _____.

The results of the sample data _____.

Do not reject the null when hypothesis when _____.

The results of the sample data _____.

Answer

Reject the null hypothesis when **the p-value is less than the established alpha value**. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when **the p-value is greater than the established alpha value**. The results of the sample data **do not support the alternative hypothesis**.

Exercise 10.5.2

It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

- $H_0 : p = 0.50, H_a : p > 0.50$
- $\alpha = 0.01$
- p-value $= 0.025$

Interpret the results and state a conclusion in simple, non-technical terms.

Answer

Since the p-value is greater than the established alpha value (the p-value is high), we do not reject the null hypothesis. There is not enough evidence to support It's a Boy Genetics Labs' stated claim that their procedures improve the chances of a boy being born.

Review

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the p-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

- $\alpha > p - value$, reject the null hypothesis
- $\alpha \leq p - value$, do not reject the null hypothesis

Glossary

Level of Significance of the Test

probability of a Type I error (reject the null hypothesis when it is true). Notation: α. In hypothesis testing, the Level of Significance is called the preconceived α or the preset α.

***p*-value**

the probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

10.6: Additional Information and Full Hypothesis Test Examples

- In a hypothesis test problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use before collecting the sample data.
- If no level of significance is given, a common standard to use is $\alpha = 0.05$.
- When you calculate the p-value and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The alternative hypothesis, H_a, tells you if the test is left, right, or two-tailed. It is the key to conducting the appropriate test.
- H_a never has a symbol that contains an equal sign.
- Thinking about the meaning of the p-value: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p-value such as 0.4, as opposed to a p-value of 0.056 ($\alpha = 0.05$ is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

Example 10.6.1

$H_0 : \mu = 5, H_a : \mu < 5$

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p-value is as follows:

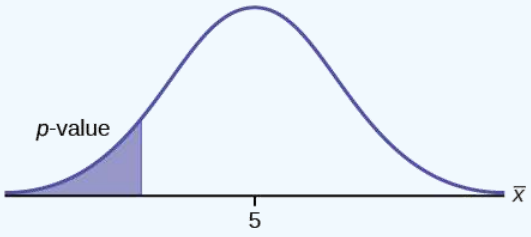


Figure 10.6.1

Exercise 10.6.1

$H_0 : \mu = 10, H_a : \mu < 10$

Assume the p-value is 0.0935. What type of test is this? Draw the picture of the p-value.

Answer

left-tailed test

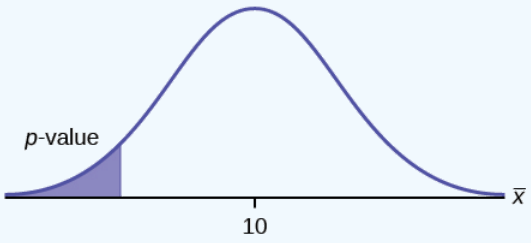


Figure 10.6.2

Example 10.6.2

$H_0 : \mu \leq 0.2, H_a : \mu > 0.2$

This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the p-value is as follows:

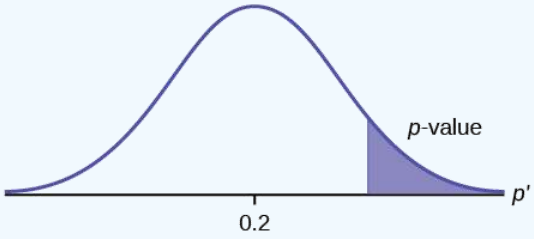


Figure 10.6.3

? Exercise 10.6.2

$H_0 : \mu \leq 1, H_a : \mu > 1$

Assume the p-value is 0.1243. What type of test is this? Draw the picture of the p-value.

Answer

right-tailed test

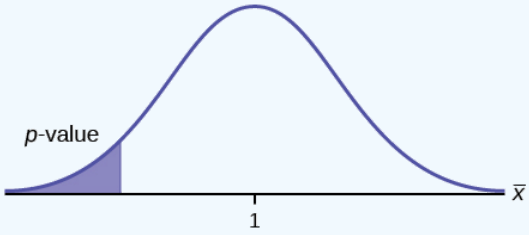


Figure 10.6.4

✓ Example 10.6.3

$H_0 : \mu = 50, H_a : \mu \neq 50$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the p-value is as follows.

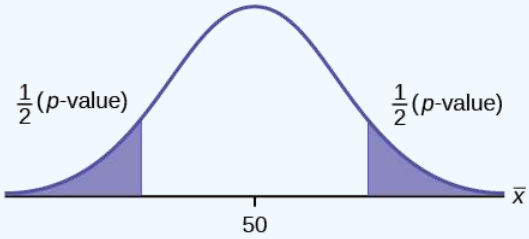


Figure 10.6.5

? Exercise 10.6.3

$H_0 : \mu = 0.5, H_a : \mu \neq 0.5$

Assume the p-value is 0.2564. What type of test is this? Draw the picture of the p-value.

Answer

two-tailed test

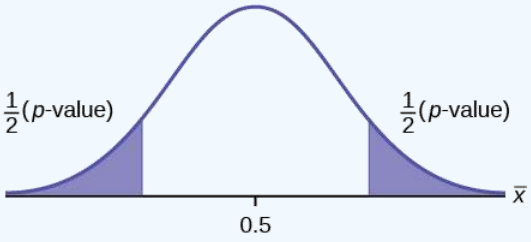


Figure 10.6.6

Full Hypothesis Test Examples

Example 10.6.4

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.**Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

$H_0 : \mu = 16.43, H_a : \mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X} =$ the mean time to swim the 25-yard freestyle.

Distribution for the test: $\bar{X}$ is normal (population standard deviation is known: $\sigma = 0.8$)

$\bar{X} - N\left(\mu, \frac{\sigma_x}{\sqrt{n}}\right)$ Therefore, $\bar{X} - N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Calculate the $p - \text{value}$ using the normal distribution for a mean:

p-value $= P(\bar{x} < 16) = 0.0187$ where the sample mean in the problem is given as 16.

p-value $= 0.0187$ (This is called the **actual level of significance**.) The $p - \text{value}$ is the area to the left of the sample mean is given as 16.

Graph:

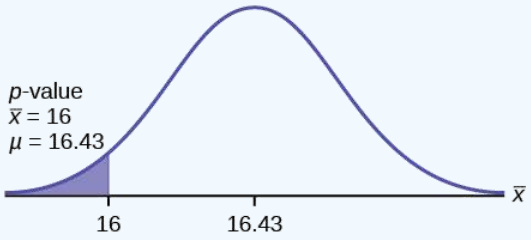


Figure 10.6.7

$\mu = 16.43$ comes from H_0. Our assumption is $\mu = 16.43$.

Interpretation of the $p - \text{value}$: If H_0 is true, there is a 0.0187 probability (1.87%) that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the $p-$value:

$\alpha = 0.05 p\text{-value} = 0.0187 \alpha > p\text{-value}$

Make a decision: Since $\alpha > p$-value, reject H_0.

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The *p*-value can easily be calculated.

Press `STAT` and arrow over to `TESTS` . Press `1:Z-Test` . Arrow over to `Stats` and press `ENTER` . Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for n. Arrow down to μ : (alternate hypothesis) and arrow over to $< \mu_0$. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The calculator not only calculates the *p*-value ($p = 0.0187$) but it also calculates the test statistic (*z*-score) for the sample mean. $\mu < 16.43$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER` . A shaded graph appears with $z = -2.08$ (test statistic) and $p = 0.0187$ ($p-$value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

When the calculator does a Z-Test, the `Z-Test` function finds the *p*-value by doing a normal probability calculation using the central limit theorem:

$P(\bar{X} < 16)$ `2nd DISTR normcdf` $((-10^{99}, 16, 16.43, \frac{0.8}{\sqrt{15}})$

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Exercise 10.6.4

The mean throwing distance of a football for a Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

Press STAT and arrow over to TESTS. Press 1: Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 40 for μ_0 (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for n. Arrow down to μ: (alternative hypothesis) and set it either as $<$, $\neq$, or $>$. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the *p*-value but it also calculates the test statistic (*z*-score) for the sample mean. Select $<$, $\neq$, or $>$ for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistic and *p*-value. Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Answer

Since the problem is about a mean, this is a test of a single population mean.

- $H_0 : \mu = 40$
- $H_a : \mu > 40$
- $p = 0.0062$

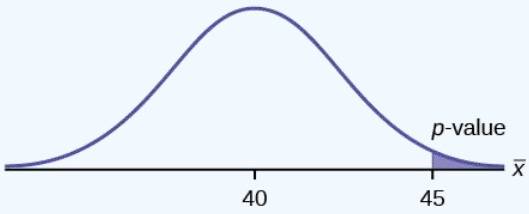


Figure 10.6.8

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the change in grip improved Marco's throwing distance.

Historical Note

The traditional way to compare the two probabilities, α and the $p-\text{value}$, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is –2.08. (From the Central Limit Theorem, the test statistic formula is $z = \frac{\bar{x}-\mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$. For this problem, $\bar{x} = 16$, $\mu_x = 16.43$ from the null hypotheses is, $\sigma_x = 0.8$, and $n = 15$.)

You can find the critical value for $\alpha = 0.05$ in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that $\alpha > p-\text{value}$), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the $p-\text{value}$, 0.0187 is considerably smaller than $\alpha = 0.05$. You can be confident about your decision to reject. The graph shows α, the $p-\text{value}$, and the test statistics and the critical value.

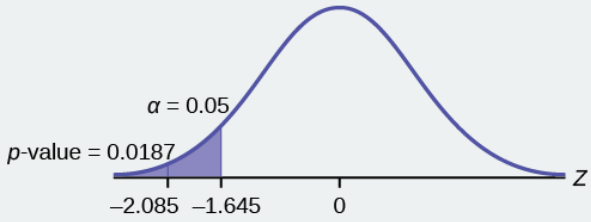


Figure 10.6.9

Example 10.6.5

A college football coach thought that his players could bench press a **mean weight of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3); 215(3); 225(1); 241(2); 252(2); 265(2); 275(2); 313(2); 316(5); 338(2); 341(1); 345(2); 368(2); 385(1).

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is more than 275 pounds.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a test of a single population mean.

- $H_0 : \mu = 275$
- $H_a : \mu > 275$

This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\bar{X} =$ the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

- $\bar{X} - N\left(275, \frac{55}{\sqrt{30}}\right)$
- $\bar{x} = 286.2$ pounds (from the data).
- $\sigma = 55$ pounds **(Always use σ if you know it.)** We assume $\mu = 275$ pounds unless our data shows us otherwise.

Calculate the *p*-value using the normal distribution for a mean and using the sample mean as input (see [link] for using the data as input):

$$p\text{-value} = P(\bar{x} > 286.2) = 0.1323.$$

Interpretation of the *p*-value: If H_0 is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.

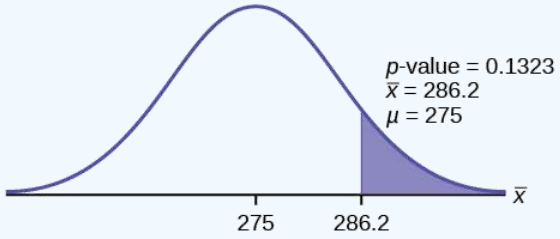


Figure 10.6.10

Compare α and the $p-\text{value}$:

$\alpha = 0.025 p - value = 0.1323$

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The $p-\text{value}$ can easily be calculated.

Put the data and frequencies into lists. Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ: and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the $p-\text{value}$ ($p = 0.1331$), a little different from the previous calculation - in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (*z*-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ ($p-\text{value}$). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

Example 10.6.6

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65 65 70 67 66 63 63 68 72 71. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Answer

Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

$H_0 : \mu = 65 \qquad H_a : \mu > 65$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: $\bar{X} =$ average score on the first statistics test.

Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t.

Use t_{df}. Therefore, the distribution for the test is t_9 where $n = 10$ and $df = 10 - 1 = 9$.

Calculate the p-value using the Student's t-distribution:

p-value $= P(\bar{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the *p*-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.

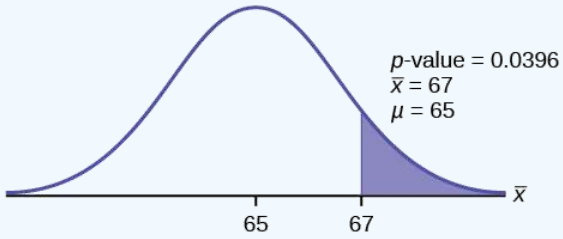


Figure 10.6.11

Compare α and the $p-$value:

Since $\alpha = 0.05$ and p-value $= 0.0396$. $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated.

Put the data into a list. Press `STAT` and arrow over to `TESTS` . Press `2:T-Test` . Arrow over to `Data` and press `ENTER` . Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for `Freq:` . Arrow down to μ: and arrow over to $> \mu_0$. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The calculator not only calculates the p-value (p = 0.0396) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER` . A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

? Exercise 10.6.6

It is believed that a stock price for a particular company will grow at a rate of \$5 per week with a standard deviation of \$1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: \$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : \mu = 5$
- $H_a : \mu < 5$
- $p = 0.0082$

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the stock price of the company grows at a rate less than \$5 a week.

- Type I Error: To conclude that the stock price is growing slower than \$5 a week when, in fact, the stock price is growing at \$5 a week (reject the null hypothesis when the null hypothesis is true).

- Type II Error: To conclude that the stock price is growing at a rate of $5 a week when, in fact, the stock price is growing slower than $5 a week (do not reject the null hypothesis when the null hypothesis is false).

Example 10.6.7

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Answer

Set up the hypothesis test:

The 1% level of significance means that $\alpha = 0.01$. This is a **test of a single population proportion**.

$H_0 : p = 0.50 \qquad H_a : p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: $P' =$ the percent of of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P', the estimated proportion.

$$P' - N\left(p, \sqrt{\frac{p-q}{n}}\right)$$

Therefore,

$$P' - N\left(0.5, \sqrt{\frac{0.5-0.5}{100}}\right)$$

where $p = 0.50, q = 1 - p = 0.50$, and $n = 100$

Calculate the p-value using the normal distribution for proportions:

$$p\text{-value} = P(p' < 0.47 \text{ or } p' > 0.53) = 0.5485$$

where

$$x = 53, p' = \frac{x}{n} = \frac{53}{100} = 0.53$$

.

Interpretation of the p-value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in Figure).

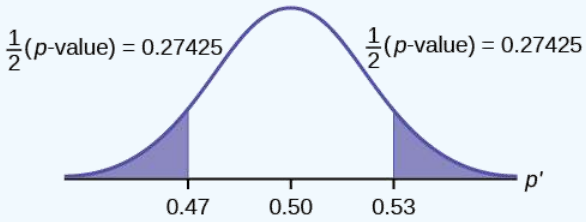


Figure 10.6.12

$\mu = p = 0.50$ comes from H_0, the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50 - 0.03 = 0.47$where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and p-value $= 0.5485$. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, you cannot reject H_0.

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p-value can easily be calculated.

Press `STAT` and arrow over to `TESTS` . Press `5:1-PropZTest` . Enter .5 for p_0, 53 for x and 100 for n. Arrow down to `Prop` and arrow to `not equals` p_0. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). `Prop not equals` .5 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER` . A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides who are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Exercise 10.6.7

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

Answer

Since the problem is about percentages, this is a test of single population proportions.

- $H_0 : p = 0.85$
- $H_a : p \neq 0.85$
- $p = 0.7554$

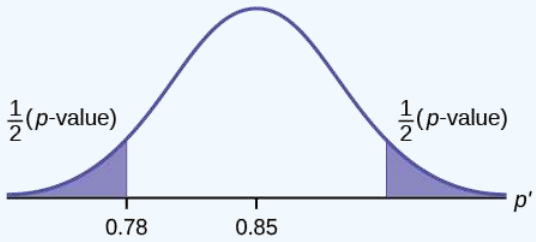


Figure 10.6.13

Because $p > \alpha$, we fail to reject the null hypothesis. There is not sufficient evidence to suggest that the proportion of students that want to go to the zoo is not 85%.

Example 10.6.8

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Answer

Set up the Hypothesis Test:

$H_0 : p = 0.30, H_a : p \neq 0.30$

Determine the distribution needed:

The **random variable** is $P' =$ proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' - N\left(0.30, \sqrt{\frac{(0.30 \cdot 0.70)}{150}}\right)$

Exercise 10.6.8.2

a. The value that helps determine the p-value is p'. Calculate p'.

Answer

a. $p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$x = 43, n = 150$

$p' = 43150$

Exercise 10.6.8.3

b. What is a **success** for this problem?

Answer

b. A success is having three cell phones in a household.

Exercise 10.6.8.4

c. What is the level of significance?

Answer

c. The level of significance is the preset α. Since α is not given, assume that $\alpha = 0.05$.

Exercise 10.6.8.5

d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.

Calculate the p-value.

Answer

d. p-value $= 0.7216$

Exercise 10.6.8.6

e. Make a decision. _____(Reject/Do not reject) H_0 because_____.

Answer

e. Assuming that $\alpha = 0.05, \alpha <$ p-value. The decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

Exercise 10.6.8

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : p = 0.92$
- $H_a : p < 0.92$
- p-value $= 0.0046$

Because $p < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

- Type I Error: To conclude that fewer than 92% of American adults own cell phones when, in fact, 92% of American adults do own cell phones (reject the null hypothesis when the null hypothesis is true).
- Type II Error: To conclude that 92% of American adults own cell phones when, in fact, fewer than 92% of American adults own cell phones (do not reject the null hypothesis when the null hypothesis is false).

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter p. The distribution for the test is normal. The estimated proportion p' is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Example 10.6.9

My dog has so many fleas,

They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25% of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!
Before his shampoo
I counted 42.

At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out

Use the new shampoo or go without?

Answer

Set up the hypothesis test:

$H_0 : p \leq 0.25 \qquad H_a : p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable $\bar{X}$ or P' represents.

$P' =$ The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal:

$$N\left(0.25, \sqrt{\frac{(0.25)1-0.25}{42}}\right)$$

Test Statistic: $z = 2.3163$

Calculate the p-value using the normal distribution for proportions:

$$p\text{-value} = 0.0103$$

In one to two complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

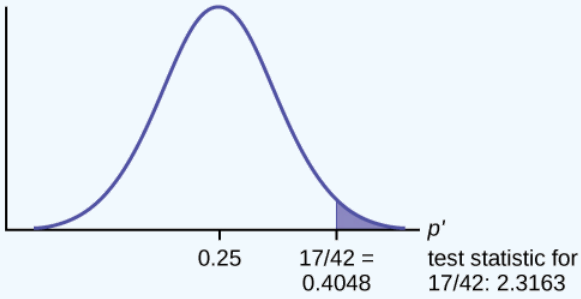


Figure 10.6.14

Compare α and the p-value:

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using complete sentences.

alpha	decision	reason for decision
0.01	Do not reject H_0	$\alpha < p$-value

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

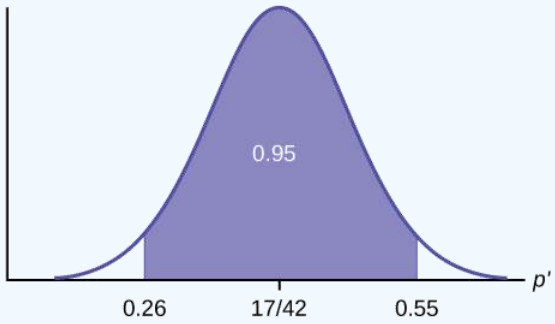


Figure 10.6.15

Confidence Interval: (0.26,0.55) We are 95% confident that the true population proportion *p* of fleas that are killed by the new shampoo is between 26% and 55%.

This test result is not very definitive since the p-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

Example 10.6.10

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95

Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

Answer

Let's follow a four-step process to answer this statistical question.

1. **State the Question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be
   a. $H_0 : \mu \leq 1$
   b. $H_a : \mu > 1$
2. **Plan**: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's-t distribution. Assume the underlying population is normal.
3. **Do the calculations**: We will input the sample data into the TI-83 as follows.

Figure 10.6.7.

Figure 10.6.8.

Figure 10.6.9.

Figure 10.6.10.

4. **State the Conclusions**: Since the p-value$(p = 0.036)$ is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Example 10.6.11

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Answer

We will follow the four-step process.

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be
   a. $H_0 : p \leq 0.00034$
   b. $H_a : p > 0.00034$

   If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.
2. We will be testing a sample proportion with $x = 172$ and $n = 420,019$. The sample is sufficiently large because we have $np = 420,019(0.00034) = 142.8 \$ nq = 420,019(0.99966) = 419,876.2$two independent outcomes, and a fixed probability of success $p = 0.00034$. Thus we will be able to generalize our results to the population.
3. The associated TI results are

   ***Figure* 10.6.11.**

   ***Figure* 10.6.12.**
4. Since the p-value $= 0.0073$ is greater than our alpha value $= 0.005$, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

Example 10.6.12

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

Answer

We will follow the four-step plan.

1. We need to test whether the proportion of sexual assaults in Daviess County, KY is significantly different from the national average.
2. Since we are presented with proportions, we will use a one-proportion *z*-test. The hypotheses for the test will be
   a. $H_0 : p = 0.00078$
   b. $H_a : p \neq 0.00078$
3. The following screen shots display the summary statistics from the hypothesis test.

   ***Figure* 10.6.13.**

   ***Figure* 10.6.14.**
4. Since the p-value, $p = 0.00063$, is less than the alpha level of 0.01, the sample data indicates that we should reject the null hypothesis. In conclusion, the sample data support the claim that the proportion of sexual assaults in Daviess County, Kentucky is different from the national average proportion.

Review

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the p-value. (A *z*-score and a *t*-score are examples of test statistics.)

5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the p-value. Remember that the quantity $1-\beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same.If the power is low, the null hypothesis might not be rejected when it should be.

Exercise 10.6.8

Assume $H_0 : \mu = 9$ and $H_a : \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a left-tailed test.

Exercise 10.6.9

Assume $H_0 : \mu \leq 6$ and $H_a : \mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise 10.6.10

Assume $H_0 : p = 0.25$ and $H_a : p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

Exercise 10.6.11

Draw the general graph of a left-tailed test.

Exercise 10.6.12

Draw the graph of a two-tailed test.

Answer

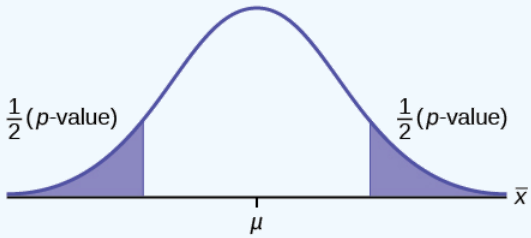


Figure 10.6.16

Exercise 10.6.13

A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

Exercise 10.6.14

Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

Answer

a right-tailed test

Exercise 10.6.15

A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

Exercise 10.6.16

You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?

Answer

a left-tailed test

Exercise 10.6.17

If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

Exercise 10.6.18

Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a left-tailed test.

Exercise 10.6.19

Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise 10.6.20

Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

References

1. Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.
2. Data from *Bloomberg Businessweek*. Available online at www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.
3. Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).
4. Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).
5. Data from *Growing by Degrees* by Allen and Seaman.
6. Data from La Leche League International. Available online at www.lalecheleague.org/Law/BAFeb01.html.
7. Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).
8. Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).
9. Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.
10. Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)
11. Data from the U.S. Census Bureau, available online at quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).
12. Data from the United States Census Bureau. Available online at www.census.gov/hhes/socdemo/language/.
13. Data from Toastmasters International. Available online at http://toastmasters.org/artisan/deta...eID=429&Page=1.
14. Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).
15. Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed

June 27, 2013).
16. “Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at research.fhda.edu/factbook/DA...t_da_2006w.pdf.
17. Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).
18. Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size n of the sample is sufficiently large, then $\bar{X} - N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X - N\left(n\mu, \sqrt{n}\sigma\right)$. If the size *n* of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

10.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Television Survey

In a recent survey, it was stated that Americans watch television on average four hours per day. Assume that $\sigma = 2$. Using your class as the sample, conduct a hypothesis test to determine if the average for students at your school is lower.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately.Shade the actual level of significance.
    a. Graph:

Figure 9.7.1.

    b. Determine the p-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Language Survey

About 42.3% of Californians and 19.6% of all Americans over age five speak a language other than English at home. Using your class as the sample, conduct a hypothesis test to determine if the percent of the students at your school who speak a language other than English at home is different from 42.3%.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____
5. Determine the test statistic using your data.

6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph:

Figure 9.7.2.

    b. Determine the p-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Jeans Survey

Suppose that young adults own an average of three pairs of jeans. Survey eight people from your class to determine if the average is higher than three. Assume the population is normal.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph:

Figure 9.7.3.

    b. Determine the p-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

10.E: Hypothesis Testing with One Sample (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

9.1: Introduction

9.2: Null and Alternative Hypotheses

Q 9.2.1

Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter $(\mu or p)$.

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

S 9.2.1

a. $H_0 : \mu = 34; H_a : \mu \neq 34$
b. $H_0 : p \leq 0.60; H_a : p > 0.60$
c. $H_0 : \mu \geq 100,000; H_a : \mu < 100,000$
d. $H_0 : p = 0.29; H_a : p \neq 0.29$
e. $H_0 : p = 0.05; H_a : p < 0.05$
f. $H_0 : \mu \leq 10; H_a : \mu > 10$
g. $H_0 : p = 0.50; H_a : p \neq 0.50$
h. $H_0 : \mu = 6; H_a : \mu \neq 6$
i. $H_0 : p \geq 0.11; H_a : p < 0.11$
j. $H_0 : \mu \leq 20,000; H_a : \mu > 20,000$

Q 9.2.2

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin? The alternative hypothesis is:

a. $p < 0.30$
b. $p \leq 0.30$
c. $p \geq 0.30$
d. $p > 0.30$

Q 9.2.3

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

a. $p = 0.20$
b. $p > 0.20$
c. $p < 0.20$
d. $p \leq 0.20$

S 9.2.3

c

Q 9.2.4

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are:

a. $H_0 : \bar{x} = 4.5, H_a : \bar{x} > 4.5$
b. $H_0 : \mu \geq 4.5, H_a : \mu < 4.5$
c. $H_0 : \mu = 4.75, H_a : \mu > 4.75$
d. $H_0 : \mu = 4.5, H_a : \mu > 4.5$

9.3: Outcomes and the Type I and Type II Errors

Q 9.3.1

State the Type I and Type II errors in complete sentences given the following statements.

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in his or her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities mean tuition cost is more than $20,000 per year.

S 9.3.1

a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
b. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%.Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more than 60% do.
c. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.
d. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
e. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more. Type II error: We conclude that 5% or more adults ride the bus to work in Los Angeles when, in fact, fewer that 5% do.
f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.
h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.

i. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in fact it is less than 11%.
j. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000. Type II error: We conclude that the average tuition cost at private universities is at most $20,000 when, in fact, it is more than $20,000.

Q 9.3.2

For statements a-j in Exercise 9.109, answer the following in complete sentences.

a. State a consequence of committing a Type I error.
b. State a consequence of committing a Type II error.

Q 9.3.3

When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II Error?

a. To conclude the drug is safe when in, fact, it is unsafe.
b. Not to conclude the drug is safe when, in fact, it is safe.
c. To conclude the drug is safe when, in fact, it is safe.
d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

S 9.3.3

b

Q 9.3.4

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is _____.

a. at least 20%, when in fact, it is less than 20%.
b. 20%, when in fact, it is 20%.
c. less than 20%, when in fact, it is at least 20%.
d. less than 20%, when in fact, it is less than 20%.

Q 9.3.4

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

a. is more than seven hours.
b. is at most seven hours.
c. is at least seven hours.
d. is less than seven hours.

S 9.3.4

d

Q 9.3.5

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test, the Type I error is:

a. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher
b. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same
c. to conclude that the mean hours per week currently is 4.5, when in fact, it is higher
d. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

9.4: Distribution Needed for Hypothesis Testing

Q 9.4.1

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is $\bar{X} \sim$ _____

a. $N\left(7.24, \frac{1.93}{\sqrt{22}}\right)$
b. $N(7.24, 1.93)$
c. t_{22}
d. t_{21}

S 9.4.1

d

9.5: Rare Events, the Sample, Decision and Conclusion

Q 9.5.1

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0 : _____ H_a : _____
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Calculate the following:
   i. $x =$ _____
   ii. $n =$ _____
   iii. $p' =$ _____
g. Calculate $\sigma_x =$ _____. Show the formula set-up.
h. State the distribution to use for the hypothesis test.
i. Find the p-value.
j. At a pre-conceived $\alpha = 0.05$, what is your:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

9.6: Additional Information and Full Hypothesis Test Examples

For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's t-distribution for one of the following homework problems, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

Q 9.6.1.

A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using $\alpha = 0.05$, is the data highly inconsistent with the claim?

S 9.6.1

a. $H_0 : \mu \geq 50,000$
b. $H_a : \mu < 50,000$
c. Let $\bar{X} =$ the average lifespan of a brand of tires.
d. normal distribution
e. $z = -2.315$
f. p-value $= 0.0103$
g. Check student's solution.
h.
   i. alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.
i. $(43,537, 49,463)$

Q 9.6.2

From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?

Q 9.6.3

The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is \$1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level?

S 9.6.3

a. $H_0 : \mu = \$1.00$
b. $H_a : \mu \neq \$1.00$
c. Let $\bar{X} =$ the average cost of a daily newspaper.
d. normal distribution
e. $z = -0.866$
f. p-value $= 0.3865$
g. Check student's solution.
h.
   i. $\alpha : 0.01$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.01.
   iv. Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is \$1. The mean cost could be \$1.
i. $(\$0.84, \$1.06)$

Q 9.6.4

An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?

Q 9.6.5

The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let $x =$ the number of sick days they took for the past year. Should the personnel team believe that the mean number is ten?

S 9.6.5

a. $H_0 : \mu = 10$
b. $H_a : \mu \neq 10$
c. Let $\bar{X}$ the mean number of sick days an employee takes per year.
d. Student's *t*-distribution
e. $t = -1.12$
f. p-value $= 0.300$
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean number of sick days is not ten.
i. $(4.9443, 11.806)$

Q 9.6.6

In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was ten. Does it appear that the mean work week has increased for women at the 5% level?

Q 9.6.7

Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?

S 9.6.7

a. $H_0 : p \geq 0.6$
b. $H_a : p < 0.6$
c. Let $P' =$ the proportion of students who feel more enriched as a result of taking Elementary Statistics.
d. normal for a single proportion
e. 1.12
f. p-value $= 0.1308$
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.
i. Confidence Interval: $(0.409, 0.654)$

   The "plus-4s" confidence interval is $(0.411, 0.648)$

Q 9.6.8

A Nissan Motor Corporation advertisement read, "The average man's I.Q. is 107. The average brown trout's I.Q. is 4. So why can't man catch brown trout?" Suppose you believe that the brown trout's mean I.Q. is greater than four. You catch 12 brown trout. A

fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief.

Q 9.6.9

Refer to Exercise 9.119. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout's mean I.Q. is **not** four.

S 9.6.9

a. $H_0 : \mu = 4$
b. $H_a : \mu \neq 4$
c. Let $\bar{X}$ the average I.Q. of a set of brown trout.
d. two-tailed Student's t-test
e. $t = 1.95$
f. p-value $= 0.076$
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05
   iv. Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.
i. $(3.8865, 5.9468)$

Q 9.6.10

According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don't believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7?

Q 9.6.11

A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the *Newsweek* poll? In complete sentences, also give three reasons why the two polls might give different results.

S 9.6.11

a. $H_0 : p \geq 0.13$
b. $H_a : p < 0.13$
c. Let $P' =$ the proportion of Americans who have seen or sensed angels
d. normal for a single proportion
e. –2.688
f. p-value $= 0.0036$
g. Check student's solution.
h.
   i. alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-valuee is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have seen or sensed an angel is less than 13%.
i. $(0, 0.0623)$

The"plus-4s" confidence interval is (0.0022, 0.0978)

Q 9.6.12

The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

Q 9.6.13

Use the “Lap time” data for Lap 4 (see [link]) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all twenty races given.

S 9.6.13

a. $H_0 : \mu \geq 129$
b. $H_a : \mu < 129$
c. Let $\bar{X} =$ the average time in seconds that Terri finishes Lap 4.
d. Student's *t*-distribution
e. $t = 1.209$
f. 0.8792
g. Check student’s solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to conclude that Terri’s mean lap time is less than 129 seconds.
i. $(128.63, 130.37)$

Q 9.6.14

Use the “Initial Public Offering” data (see [link]) to test the claim that the mean offer price was $18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

Note

The following questions were written by past students. They are excellent problems!

Q 9.6.15

"Asian Family Reunion," by Chau Nguyen

Every two years it comes around.

We all get together from different towns.

In my honest opinion,

It's not a typical family reunion.

Not forty, or fifty, or sixty,

But how about seventy companions!

The kids would play, scream, and shout

One minute they're happy, another they'll pout.

The teenagers would look, stare, and compare

From how they look to what they wear.

The men would chat about their business

That they make more, but never less.

Money is always their subject

And there's always talk of more new projects.

The women get tired from all of the chats

They head to the kitchen to set out the mats.

Some would sit and some would stand

Eating and talking with plates in their hands.

Then come the games and the songs

And suddenly, everyone gets along!

With all that laughter, it's sad to say

That it always ends in the same old way.

They hug and kiss and say "good-bye"

And then they all begin to cry!

I say that 60 percent shed their tears

But my mom counted 35 people this year.

She said that boys and men will always have their pride,

So we won't ever see them cry.

I myself don't think she's correct,

So could you please try this problem to see if you object?

S 9.6.15

a. $H_0 : p = 0.60$
b. $H_a : p < 0.60$
c. Let $P' =$ the proportion of family members who shed tears at a reunion.
d. normal for a single proportion
e. –1.71
f. 0.0438
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value $< \alpha$
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of family members who shed tears at a reunion is less than 0.60. However, the test is weak because the p-value and alpha are quite close, so other tests should be done.
i. We are 95% confident that between 38.29% and 61.71% of family members will shed tears at a family reunion. $(0.3829, 0.6171)$ The"plus-4s" confidence interval (see chapter 8) is $(0.3861, 0.6139)$

Note that here the "large-sample" $1 - \text{PropZTest}$ provides the approximate p-value of 0.0438. Whenever a p-value based on a normal approximation is close to the level of significance, the exact p-value based on binomial probabilities should be calculated whenever possible. This is beyond the scope of this course.

Q 9.6.16

"The Problem with Angels," by Cyndy Dowling

Although this problem is wholly mine,

The catalyst came from the magazine, Time.

On the magazine cover I did find

The realm of angels tickling my mind.

Inside, 69% I found to be

In angels, Americans do believe.

Then, it was time to rise to the task,

Ninety-five high school and college students I did ask.

Viewing all as one group,

Random sampling to get the scoop.

So, I asked each to be true,

"Do you believe in angels?" Tell me, do!

Hypothesizing at the start,

Totally believing in my heart

That the proportion who said yes

Would be equal on this test.

Lo and behold, seventy-three did arrive,

Out of the sample of ninety-five.

Now your job has just begun,

Solve this problem and have some fun.

Q 9.6.17

"Blowing Bubbles," by Sondra Prull

Studying stats just made me tense,

I had to find some sane defense.

Some light and lifting simple play

To float my math anxiety away.

Blowing bubbles lifts me high

Takes my troubles to the sky.

POIK! They're gone, with all my stress

Bubble therapy is the best.

The label said each time I blew

The average number of bubbles would be at least 22.

I blew and blew and this I found

From 64 blows, they all are round!

But the number of bubbles in 64 blows

Varied widely, this I know.

20 per blow became the mean

They deviated by 6, and not 16.

From counting bubbles, I sure did relax

But now I give to you your task.

Was 22 a reasonable guess?

Find the answer and pass this test!

S 9.6.17

a. $H_0 : \mu \geq 22$
b. $H_a : \mu < 22$
c. Let $\bar{X} =$ the mean number of bubbles per blow.
d. Student's *t*-distribution

e. –2.667
f. p-value $= 0.00486$
g. Check student's solution.
h. i. α : 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The p-value is less than 0.05.
iv. Conclusion: There is sufficient evidence to conclude that the mean number of bubbles per blow is less than 22.
i. $(18.501, 21.499)$

Q 9.6.18

"Dalmatian Darnation," by Kathy Sparling

A greedy dog breeder named Spreckles

Bred puppies with numerous freckles

The Dalmatians he sought

Possessed spot upon spot

The more spots, he thought, the more shekels.

His competitors did not agree

That freckles would increase the fee.

They said, “Spots are quite nice

But they don't affect price;

One should breed for improved pedigree.”

The breeders decided to prove

This strategy was a wrong move.

Breeding only for spots

Would wreak havoc, they thought.

His theory they want to disprove.

They proposed a contest to Spreckles

Comparing dog prices to freckles.

In records they looked up

One hundred one pups:

Dalmatians that fetched the most shekels.

They asked Mr. Spreckles to name

An average spot count he'd claim

To bring in big bucks.

Said Spreckles, “Well, shucks,

It's for one hundred one that I aim.”

Said an amateur statistician

Who wanted to help with this mission.

“Twenty-one for the sample

Standard deviation's ample:

They examined one hundred and one

Dalmatians that fetched a good sum.

They counted each spot,

Mark, freckle and dot

And tallied up every one.

Instead of one hundred one spots

They averaged ninety six dots

Can they muzzle Spreckles'

Obsession with freckles

Based on all the dog data they've got?

Q 9.6.19

"Macaroni and Cheese, please!!" by Nedda Misherghi and Rachelle Hall

As a poor starving student I don't have much money to spend for even the bare necessities. So my favorite and main staple food is macaroni and cheese. It's high in taste and low in cost and nutritional value.

One day, as I sat down to determine the meaning of life, I got a serious craving for this, oh, so important, food of my life. So I went down the street to Greatway to get a box of macaroni and cheese, but it was SO expensive! $2.02 !!! Can you believe it? It made me stop and think. The world is changing fast. I had thought that the mean cost of a box (the normal size, not some super-gigantic-family-value-pack) was at most $1, but now I wasn't so sure. However, I was determined to find out. I went to 53 of the closest grocery stores and surveyed the prices of macaroni and cheese. Here are the data I wrote in my notebook:

Price per box of Mac and Cheese:

- 5 stores @ $2.02
- 15 stores @ $0.25
- 3 stores @ $1.29
- 6 stores @ $0.35
- 4 stores @ $2.27
- 7 stores @ $1.50
- 5 stores @ $1.89
- 8 stores @ 0.75.

I could see that the cost varied but I had to sit down to figure out whether or not I was right. If it does turn out that this mouth-watering dish is at most $1, then I'll throw a big cheesy party in our next statistics lab, with enough macaroni and cheese for just me. (After all, as a poor starving student I can't be expected to feed our class of animals!)

S 9.6.19

a. $H_0 : \mu \leq 1$
b. $H_a : \mu > 1$
c. Let $\bar{X} =$ the mean cost in dollars of macaroni and cheese in a certain town.
d. Student's t-distribution
e. $t = 0.340$
f. p-value $= 0.36756$
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05
   iv. Conclusion: The mean cost could be $1, or less. At the 5% significance level, there is insufficient evidence to conclude that the mean price of a box of macaroni and cheese is more than $1.
i. $(0.8291, 1.241)$

Q 9.6.20

"William Shakespeare: The Tragedy of Hamlet, Prince of Denmark," by Jacqueline Ghodsi

THE CHARACTERS (in order of appearance):

- HAMLET, Prince of Denmark and student of Statistics
- POLONIUS, Hamlet's tutor
- HOROTIO, friend to Hamlet and fellow student

Scene: The great library of the castle, in which Hamlet does his lessons

Act I

(The day is fair, but the face of Hamlet is clouded. He paces the large room. His tutor, Polonius, is reprimanding Hamlet regarding the latter's recent experience. Horatio is seated at the large table at right stage.)

POLONIUS: My Lord, how cans't thou admit that thou hast seen a ghost! It is but a figment of your imagination!

HAMLET: I beg to differ; I know of a certainty that five-and-seventy in one hundred of us, condemned to the whips and scorns of time as we are, have gazed upon a spirit of health, or goblin damn'd, be their intents wicked or charitable.

POLONIUS If thou doest insist upon thy wretched vision then let me invest your time; be true to thy work and speak to me through the reason of the null and alternate hypotheses. (He turns to Horatio.) Did not Hamlet himself say, "What piece of work is man, how noble in reason, how infinite in faculties? Then let not this foolishness persist. Go, Horatio, make a survey of three-and-sixty and discover what the true proportion be. For my part, I will never succumb to this fantasy, but deem man to be devoid of all reason should thy proposal of at least five-and-seventy in one hundred hold true.

HORATIO (to Hamlet): What should we do, my Lord?

HAMLET: Go to thy purpose, Horatio.

HORATIO: To what end, my Lord?

HAMLET: That you must teach me. But let me conjure you by the rights of our fellowship, by the consonance of our youth, but the obligation of our ever-preserved love, be even and direct with me, whether I am right or no.

(Horatio exits, followed by Polonius, leaving Hamlet to ponder alone.)

Act II

(The next day, Hamlet awaits anxiously the presence of his friend, Horatio. Polonius enters and places some books upon the table just a moment before Horatio enters.)

POLONIUS: So, Horatio, what is it thou didst reveal through thy deliberations?

HORATIO: In a random survey, for which purpose thou thyself sent me forth, I did discover that one-and-forty believe fervently that the spirits of the dead walk with us. Before my God, I might not this believe, without the sensible and true avouch of mine own eyes.

POLONIUS: Give thine own thoughts no tongue, Horatio. (Polonius turns to Hamlet.) But look to't I charge you, my Lord. Come Horatio, let us go together, for this is not our test. (Horatio and Polonius leave together.)

HAMLET: To reject, or not reject, that is the question: whether 'tis nobler in the mind to suffer the slings and arrows of outrageous statistics, or to take arms against a sea of data, and, by opposing, end them. (Hamlet resignedly attends to his task.)

(Curtain falls)

Q 9.6.21

"Untitled," by Stephen Chen

I've often wondered how software is released and sold to the public. Ironically, I work for a company that sells products with known problems. Unfortunately, most of the problems are difficult to create, which makes them difficult to fix. I usually use the test program X, which tests the product, to try to create a specific problem. When the test program is run to make an error occur, the likelihood of generating an error is 1%.

So, armed with this knowledge, I wrote a new test program Y that will generate the same error that test program X creates, but more often. To find out if my test program is better than the original, so that I can convince the management that I'm right, I ran my test program to find out how often I can generate the same error. When I ran my test program 50 times, I generated the error twice. While this may not seem much better, I think that I can convince the management to use my test program instead of the original test program. Am I right?

S 9.6.21

a. $H_0 : p = 0.01$
b. $H_a : p > 0.01$
c. Let $P' =$ the proportion of errors generated
d. Normal for a single proportion
e. 2.13
f. 0.0165
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of errors generated is more than 0.01.
i. Confidence interval: $(0, 0.094)$

The"plus-4s" confidence interval is $(0.004, 0.144)$

Q 9.6.22

"Japanese Girls' Names"

by Kumi Furuichi

It used to be very typical for Japanese girls' names to end with "ko." (The trend might have started around my grandmothers' generation and its peak might have been around my mother's generation.) "Ko" means "child" in Chinese characters. Parents would name their daughters with "ko" attaching to other Chinese characters which have meanings that they want their daughters to become, such as Sachiko—happy child, Yoshiko—a good child, Yasuko—a healthy child, and so on.

However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names which end with "ko." More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children.

I have a feeling that, while 70 percent or more of my mother's generation would have names with "ko" at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends', ex-classmates', co-workers, and acquaintances' names that I could remember. Following are the names. (Some are repeats.) Test to see if the proportion has dropped for this generation.

Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi, Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.

Q 9.6.23

"Phillip's Wish," by Suzanne Osorio

My nephew likes to play

Chasing the girls makes his day.

He asked his mother

If it is okay

To get his ear pierced.

She said, “No way!”

To poke a hole through your ear,

Is not what I want for you, dear.

He argued his point quite well,

Says even my macho pal, Mel,

Has gotten this done.

It’s all just for fun.

C’mon please, mom, please, what the hell.

Again Phillip complained to his mother,

Saying half his friends (including their brothers)

Are piercing their ears

And they have no fears

He wants to be like the others.

She said, “I think it’s much less.

We must do a hypothesis test.

And if you are right,

I won’t put up a fight.

But, if not, then my case will rest.”

We proceeded to call fifty guys

To see whose prediction would fly.

Nineteen of the fifty

Said piercing was nifty

And earrings they’d occasionally buy.

Then there’s the other thirty-one,

Who said they’d never have this done.

So now this poem’s finished.

Will his hopes be diminished,

Or will my nephew have his fun?

S 9.6.23

a. $H_0 : p = 0.50$
b. $H_a : p < 0.50$
c. Let P' = the proportion of friends that has a pierced ear.
d. normal for a single proportion
e. –1.70
f. p-value $= 0.0448$
g. Check student’s solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: The p-value is less than 0.05. (However, they are very close.)
   iv. Conclusion: There is sufficient evidence to support the claim that less than 50% of his friends have pierced ears.
i. Confidence Interval: $(0.245, 0.515)$ The “plus-4s” confidence interval is $(0.259, 0.519)$

Q 9.6.24

"The Craven," by Mark Salangsang

Once upon a morning dreary

In stats class I was weak and weary.

Pondering over last night's homework

Whose answers were now on the board

This I did and nothing more.

While I nodded nearly napping

Suddenly, there came a tapping.

As someone gently rapping,

Rapping my head as I snore.

Quoth the teacher, "Sleep no more."

"In every class you fall asleep,"

The teacher said, his voice was deep.

"So a tally I've begun to keep

Of every class you nap and snore.

The percentage being forty-four."

"My dear teacher I must confess,

While sleeping is what I do best.

The percentage, I think, must be less,

A percentage less than forty-four."

This I said and nothing more.

"We'll see," he said and walked away,

And fifty classes from that day

He counted till the month of May

The classes in which I napped and snored.

The number he found was twenty-four.

At a significance level of 0.05,

Please tell me am I still alive?

Or did my grade just take a dive

Plunging down beneath the floor?

Upon thee I hereby implore.

Q 9.6.25

Toastmasters International cites a report by Gallop Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percent at her school is less than 40%.

S 9.6.25

a. $H_0 : p = 0.40$
b. $H_a : p < 0.40$

c. Let P' = the proportion of schoolmates who fear public speaking.
d. normal for a single proportion
e. –1.01
f. p-value $= 0.1563$
g. Check student's solution.
h. i. α : 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The p-value is greater than 0.05.
iv. Conclusion: There is insufficient evidence to support the claim that less than 40% of students at the school fear public speaking.
i. Confidence Interval: $(0.3241, 0.4240)$; The "plus-4s" confidence interval is $(0.3257, 0.4250)$

Q 9.6.26

Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California's percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California, was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68% represents California. NOTE: For more accurate results, use more California community colleges and this past year's data.

Q 9.6.27

According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased.

S 9.6.27

a. $H_0 : p = 0.14$
b. $H_a : p < 0.14$
c. Let P' = the proportion of NYC residents that smoke.
d. normal for a single proportion
e. –0.2756
f. p-value $= 0.3914$
g. Check student's solution.
h. i. α : 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The p-value is greater than 0.05.
iv. At the 5% significance level, there is insufficient evidence to conclude that the proportion of NYC residents who smoke is less than 0.14.
i. Confidence Interval: $(0.0502, 0.2070)$; The "plus-4s" confidence interval (see chapter 8) is $(0.0676, 0.2297)$

Q 9.6.28

The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test.

Q 9.6.29

Registered nurses earned an average annual salary of \$69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than \$69,110 for California nurses. The sample average was \$71,121 with a sample standard deviation of \$7,489. Conduct a hypothesis test.

S 9.6.29

a. $H_0 : \mu = 69,110$
b. $H_0 : \mu > 69,110$
c. Let $\bar{X}$ = the mean salary in dollars for California registered nurses.

d. Student's t-distribution
e. $t = 1.719$
f. p-value : 0.0466
g. Check student's solution.
h.
   i. α : 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.
i. $(\$68,757, \$73,485)$

Q 9.6.30

La Leche League International reports that the mean age of weaning a child from breastfeeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the U.S. is less than four years old.

Q 9.6.31

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin?

After conducting the test, your decision and conclusion are

a. Reject H_0: There is sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
b. Do not reject H_0: There is not sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.
c. Do not reject H_0: There is not sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
d. Reject H_0: There is sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.

S 9.6.31

c

Q 9.6.32

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing.

At a 1% level of significance, an appropriate conclusion is:

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20%.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20%.

Q 9.6.33

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.

At a significance level of $a = 0.05$, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75
b. There is enough evidence to conclude that the mean number of hours is more than 4.5
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75

S 9.6.33

c

Instructions: For the following ten exercises,

Hypothesis testing: For the following ten exercises, answer each question.

State the null and alternate hypothesis.

State the p-value.

State α.

What is your decision?

Write a conclusion.

Answer any other questions asked in the problem.

Q 9.6.34

According to the Center for Disease Control website, in 2011 at least 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized–approximately 1,200 students–small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of the 150 students surveyed, 82 have smoked. Use a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

Q 9.6.35

A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

S 9.6.35

a. $H_0 : p = 0.488$ $H_a : p \neq 0.488$
b. p-value $= 0.0114$
c. $\alpha = 0.05$
d. Reject the null hypothesis.
e. At the 5% level of significance, there is enough evidence to conclude that 48.8% of families own stocks.
f. The survey does not appear to be accurate.

Q 9.6.36

Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

Q 9.6.37

The US Department of Energy reported that 51.7% of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level in Kentucky? Are the results applicable across the country? Why?

S 9.6.37

a. $H_0 : p = 0.517$ $H_0 : p \neq 0.517$
b. p-value $= 0.9203$.
c. $\alpha = 0.05$.
d. Do not reject the null hypothesis.

e. At the 5% significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.
f. However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

Q 9.6.38

For Americans using library services, the American Library Association claims that at most 67% of patrons borrow books. The library director in Owensboro, Kentucky feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

Q 9.6.39

The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05 level$, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

S 9.6.39

a. $H_0 : \mu \geq 11.52$ $H_a : \mu < 11.52$
b. p-value $= 0.000002$which is almost 0.
c. $\alpha = 0.05$.
d. Reject the null hypothesis.
e. At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeaster US is less than 11.52 inches, on average.
f. We would make the same conclusion if alpha was 1% because the p-value is almost 0.

Q 9.6.40

A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

Q 9.6.41

A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals

3; 2; 1; 3; 7; 2; 9; 4; 6; 6; 8; 0; 5; 6; 4; 2; 1; 3; 4; 1

At the $\alpha = 0.05$ level can it be concluded that the sample mean is higher than 5.8 visits per year?

S 9.6.42

1. $H_0 : \mu \leq 5.8$ $H_a : \mu > 5.8$
2. p-value $= 0.9987$
3. $\alpha = 0.05$
4. Do not reject the null hypothesis.
5. At the 5% level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

Q 9.6.42

According to the *N.Y. Times Almanac* the mean family size in the U.S. is 3.18. A sample of a college math class resulted in the following family sizes:

5; 4; 5; 4; 4; 3; 6; 4; 3; 3; 5; 5; 6; 3; 3; 2; 7; 4; 5; 2; 2; 2; 3; 2

At $\alpha = 0.05$ level, is the class' mean family size greater than the national average? Does the Almanac result remain valid? Why?

Q 9.6.43

The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is the student academic group's claim correct?

S 9.6.43

a. $H_0 : \mu \geq 150$ $H_0 : \mu < 150$
b. p-value $= 0.0622$
c. $\alpha = 0.01$
d. Do not reject the null hypothesis.
e. At the 1% significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average.
f. The student academic group's claim appears to be correct.

9.7: Hypothesis Testing of a Single Mean and Single Proportion

Index

A

Adding probabilities
- 5.2: Complements, Intersections, and Unions

B

bar graph
- 2.3: Stem-and-Leaf Displays

binomial probability distribution
- 6.3: The Binomial Distribution
- 9.4: A Population Proportion

binomial random variable
- 6.3: The Binomial Distribution

box plots
- 2.6: Box Plots

C

central limit theorem
- 8.2: The Sampling Distribution of the Sample Mean
- 8.4: Using the Central Limit Theorem

cluster sampling
- 1.10: Methods of Sampling

complement
- 5.2: Complements, Intersections, and Unions

conditional probability
- 5.3: Conditional Probability and Independent Events

Confidence Interval
- 10.1: Prelude to Hypothesis Testing

continuous data
- 1.10: Methods of Sampling

Cumulative Normal Probability
- 7.2: The Standard Normal Distribution

cumulative probability distributions
- 6.3: The Binomial Distribution

D

Decision
- 10.5: Rare Events, the Sample, Decision and Conclusion

DENSITY FUNCTION
- 7.1: Continuous Random Variables

discrete data
- 1.10: Methods of Sampling

E

ELEMENT
- 5.1: Sample Spaces, Events, and Their Probabilities

F

Frequency Polygons
- 2.2: Organizing and Graphing Quantitative Data

H

Histograms
- 2.2: Organizing and Graphing Quantitative Data

hypothesis testing
- 10.1: Prelude to Hypothesis Testing
- 10.2: Null and Alternative Hypotheses
- 10.4: Distribution Needed for Hypothesis Testing
- 10.6: Additional Information and Full Hypothesis Test Examples

I

independent events
- 5.3: Conditional Probability and Independent Events

inferential statistics
- 9.1: Prelude to Confidence Intervals

INTERSECTIONS
- 5.2: Complements, Intersections, and Unions

J

joint probabilities
- 4.5: Two-Way Tables (3 of 5)

L

Law of Large Numbers
- 8.4: Using the Central Limit Theorem

line graph
- 2.3: Stem-and-Leaf Displays

M

margin of error
- 9.2: A Single Population Mean using the Normal Distribution

mean
- 2.8: Skewness and the Mean, Median, and Mode
- 6.2: Probability Distributions for Discrete Random Variables

mean of the sample proportion
- 8.3: The Sample Proportion

median
- 2.4: Measures of Central Tendency- Mean, Median and Mode
- 2.5: Measures of Position- Percentiles and Quartiles
- 2.8: Skewness and the Mean, Median, and Mode

mode
- 2.4: Measures of Central Tendency- Mean, Median and Mode
- 2.8: Skewness and the Mean, Median, and Mode

mutually exclusive
- 5.2: Complements, Intersections, and Unions

N

Normal Approximation to the Binomial Distribution
- 8.4: Using the Central Limit Theorem

normal distribution
- 7.1: Continuous Random Variables

O

OCCURRENCE
- 5.1: Sample Spaces, Events, and Their Probabilities

outliers
- 2.5: Measures of Position- Percentiles and Quartiles

P

Pareto chart
- 1.10: Methods of Sampling

population mean
- 2.4: Measures of Central Tendency- Mean, Median and Mode

Population Standard Deviation
- 2.7: Measures of Spread- Variance and Standard Deviation

power of the test
- 10.3: Outcomes and the Type I and Type II Errors
- 10.6: Additional Information and Full Hypothesis Test Examples

probability distribution function
- 6.2: Probability Distributions for Discrete Random Variables

Probability Rule for Complements
- 5.2: Complements, Intersections, and Unions

Q

Qualitative Data
- 1.10: Methods of Sampling

Quantitative Data
- 1.10: Methods of Sampling

quartiles
- 2.5: Measures of Position- Percentiles and Quartiles

R

random experiment
- 5.1: Sample Spaces, Events, and Their Probabilities

rare events
- 10.5: Rare Events, the Sample, Decision and Conclusion

S

sample mean
- 2.4: Measures of Central Tendency- Mean, Median and Mode
- 8.1: The Mean and Standard Deviation of the Sample Mean

sample proportion
- 8.3: The Sample Proportion

sample space
- 5.1: Sample Spaces, Events, and Their Probabilities

sample Standard Deviation
- 2.7: Measures of Spread- Variance and Standard Deviation
- 8.1: The Mean and Standard Deviation of the Sample Mean

Sampling Bias
- 1.10: Methods of Sampling

sampling distribution
- 8.3: The Sample Proportion

Sampling Error
- 1.10: Methods of Sampling

sampling with replacement
- 1.10: Methods of Sampling

sampling without replacement
- 1.10: Methods of Sampling

significance level
- 10.5: Rare Events, the Sample, Decision and Conclusion

Skewed
- 2.6: Box Plots
- 2.8: Skewness and the Mean, Median, and Mode

SPECIFICITY OF A DIAGNOSTIC TEST
- 5.3: Conditional Probability and Independent Events

standard deviation
- 2.7: Measures of Spread- Variance and Standard Deviation
- 6.2: Probability Distributions for Discrete Random Variables

standard deviation of the sample proportion
8.3: The Sample Proportion
standard normal random variable
7.2: The Standard Normal Distribution
stemplot
2.3: Stem-and-Leaf Displays

T

tails
7.4: Areas of Tails of Distributions
The alternative hypothesis
10.2: Null and Alternative Hypotheses
The null hypothesis
10.2: Null and Alternative Hypotheses
Time Series Graphs
2.2: Organizing and Graphing Quantitative Data
tree diagram
5.1: Sample Spaces, Events, and Their Probabilities
type I error
10.3: Outcomes and the Type I and Type II Errors
type II error
10.3: Outcomes and the Type I and Type II Errors

U

unions
5.2: Complements, Intersections, and Unions

V

Venn diagram
5.1: Sample Spaces, Events, and Their Probabilities

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: MA336: Statistics

Webpages: 116

Applicable Restrictions: Noncommercial

All licenses found:

- Undeclared: 55.2% (64 pages)
- CC BY 4.0: 26.7% (31 pages)
- CC BY-NC-SA 4.0: 18.1% (21 pages)

By Page

- MA336: Statistics - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Introduction to Statistical Studies - *Undeclared*
    - 1.1: Why It Matters- Types of Statistical Studies and Producing Data - *Undeclared*
    - 1.2: Introduction to Types of Statistical Studies - *Undeclared*
    - 1.3: Types of Statistical Studies (1 of 4) - *Undeclared*
    - 1.4: Types of Statistical Studies (2 of 4) - *Undeclared*
    - 1.5: Types of Statistical Studies (3 of 4) - *Undeclared*
    - 1.6: Types of Statistical Studies (4 of 4) - *Undeclared*
    - 1.7: Introduction to Sampling - *Undeclared*
    - 1.8: Sampling (1 of 2) - *Undeclared*
    - 1.9: Sampling (2 of 2) - *Undeclared*
    - 1.10: Methods of Sampling - *CC BY 4.0*
    - 1.11: Introduction to Conducting Experiments - *Undeclared*
    - 1.12: Conducting Experiments (1 of 2) - *Undeclared*
    - 1.13: Conducting Experiments (2 of 2) - *Undeclared*
    - 1.14: Putting It Together- Types of Statistical Studies and Producing Data - *Undeclared*
  - 2: Descriptive Statistics - *CC BY 4.0*
    - 2.1: Organizing and Graphing Qualitative Data - *CC BY 4.0*
    - 2.2: Organizing and Graphing Quantitative Data - *CC BY 4.0*
    - 2.3: Stem-and-Leaf Displays - *CC BY 4.0*
    - 2.4: Measures of Central Tendency- Mean, Median and Mode - *CC BY 4.0*
    - 2.5: Measures of Position- Percentiles and Quartiles - *CC BY 4.0*
    - 2.6: Box Plots - *CC BY 4.0*
    - 2.7: Measures of Spread- Variance and Standard Deviation - *CC BY 4.0*
    - 2.8: Skewness and the Mean, Median, and Mode - *CC BY 4.0*
  - 3: Examining Relationships- Quantitative Data - *Undeclared*
    - 3.1: Why It Matters- Examining Relationships- Quantitative Data - *Undeclared*
    - 3.2: Linear Regression (4 of 4) - *Undeclared*
    - 3.3: Introduction to Assessing the Fit of a Line - *Undeclared*
    - 3.4: Assessing the Fit of a Line (1 of 4) - *Undeclared*
    - 3.5: Assessing the Fit of a Line (2 of 4) - *Undeclared*
    - 3.6: Assessing the Fit of a Line (3 of 4) - *Undeclared*
    - 3.7: Assessing the Fit of a Line (4 of 4) - *Undeclared*
    - 3.8: Putting It Together- Examining Relationships- Quantitative Data - *Undeclared*
    - 3.9: StatTutor- Academic Performance - *Undeclared*
    - 3.10: Assignment- Scatterplot - *Undeclared*
    - 3.11: Assignment- Linear Relationships - *Undeclared*
    - 3.12: Introduction to Scatterplots - *Undeclared*
    - 3.13: Assignment- Linear Regression - *Undeclared*
    - 3.14: Scatterplots (1 of 5) - *Undeclared*
    - 3.15: Scatterplots (2 of 5) - *Undeclared*
    - 3.16: Scatterplots (3 of 5) - *Undeclared*
    - 3.17: Scatterplots (4 of 5) - *Undeclared*
    - 3.18: Scatterplots (5 of 5) - *Undeclared*
    - 3.19: Introduction to Linear Relationships - *Undeclared*
    - 3.20: Linear Relationships (1 of 4) - *Undeclared*
    - 3.21: Linear Relationships (2 of 4) - *Undeclared*
    - 3.22: Linear Relationships (3 of 4) - *Undeclared*
    - 3.23: Linear Relationships (4 of 4) - *Undeclared*
    - 3.24: Introduction to Association vs Causation - *Undeclared*
    - 3.25: Causation and Lurking Variables (1 of 2) - *Undeclared*

- 3.26: Causation and Lurking Variables (2 of 2) - *Undeclared*
- 3.27: Introduction to Linear Regression - *Undeclared*
- 3.28: Linear Regression (1 of 4) - *Undeclared*
- 3.29: Linear Regression (2 of 4) - *Undeclared*
- 3.30: Linear Regression (3 of 4) - *Undeclared*

- 4: Relationships in Categorical Data with Intro to Probability - *Undeclared*
  - 4.1: Why It Matters- Relationships in Categorical Data with Intro to Probability - *Undeclared*
  - 4.2: Introduction to Two-Way Tables - *Undeclared*
  - 4.3: Two-Way Tables (1 of 5) - *Undeclared*
  - 4.4: Two-Way Tables (2 of 5) - *Undeclared*
  - 4.5: Two-Way Tables (3 of 5) - *Undeclared*
  - 4.6: Two-Way Tables (4 of 5) - *Undeclared*
  - 4.7: Two-Way Tables (5 of 5) - *Undeclared*
  - 4.8: Putting It Together- Relationships in Categorical Data with Intro to Probability - *Undeclared*
- 5: Basic Concepts of Probability - *CC BY-NC-SA 4.0*
  - 5.1: Sample Spaces, Events, and Their Probabilities - *CC BY-NC-SA 4.0*
  - 5.2: Complements, Intersections, and Unions - *CC BY-NC-SA 4.0*
  - 5.3: Conditional Probability and Independent Events - *CC BY-NC-SA 4.0*
  - 5.E: Basic Concepts of Probability (Exercises) - *CC BY-NC-SA 4.0*
- 6: Discrete Random Variables - *CC BY-NC-SA 4.0*
  - 6.1: Random Variables - *CC BY-NC-SA 4.0*
  - 6.2: Probability Distributions for Discrete Random Variables - *CC BY-NC-SA 4.0*
  - 6.3: The Binomial Distribution - *CC BY-NC-SA 4.0*
  - 6.E: Discrete Random Variables (Exercises) - *CC BY-NC-SA 4.0*
- 7: Continuous Random Variables - *CC BY-NC-SA 4.0*
  - 7.1: Continuous Random Variables - *CC BY-NC-SA 4.0*
  - 7.2: The Standard Normal Distribution - *CC BY-NC-SA 4.0*
  - 7.3: Probability Computations for General Normal Random Variables - *CC BY-NC-SA 4.0*
  - 7.4: Areas of Tails of Distributions - *CC BY-NC-SA 4.0*
  - 7.E: Continuous Random Variables (Exercises) - *CC BY-NC-SA 4.0*
- 8: Sampling Distributions - *CC BY-NC-SA 4.0*
  - 8.1: The Mean and Standard Deviation of the Sample Mean - *CC BY-NC-SA 4.0*
  - 8.2: The Sampling Distribution of the Sample Mean - *CC BY-NC-SA 4.0*
  - 8.3: The Sample Proportion - *CC BY-NC-SA 4.0*
  - 8.4: Using the Central Limit Theorem - *CC BY 4.0*
    - 8.4E: Using the Central Limit Theorem (Exercises) - *CC BY 4.0*
  - 8.E: Sampling Distributions (Exercises) - *CC BY-NC-SA 4.0*
- 9: Confidence Intervals - *CC BY 4.0*
  - 9.1: Prelude to Confidence Intervals - *CC BY 4.0*
  - 9.2: A Single Population Mean using the Normal Distribution - *CC BY 4.0*
  - 9.3: A Single Population Mean using the Student t-Distribution - *CC BY 4.0*
  - 9.4: A Population Proportion - *CC BY 4.0*
  - 9.5: Confidence Interval - Home Costs (Worksheet) - *CC BY 4.0*
  - 9.6: Confidence Interval -Place of Birth (Worksheet) - *CC BY 4.0*
  - 9.7: Confidence Interval -Women's Heights (Worksheet) - *CC BY 4.0*
  - 9.E: Confidence Intervals (Exercises) - *CC BY 4.0*
  - 9.S: Confidence Intervals (Summary) - *CC BY 4.0*
- 10: Hypothesis Testing with One Sample - *CC BY 4.0*
  - 10.1: Prelude to Hypothesis Testing - *CC BY 4.0*
  - 10.2: Null and Alternative Hypotheses - *CC BY 4.0*
  - 10.3: Outcomes and the Type I and Type II Errors - *CC BY 4.0*
  - 10.4: Distribution Needed for Hypothesis Testing - *CC BY 4.0*
  - 10.5: Rare Events, the Sample, Decision and Conclusion - *CC BY 4.0*
  - 10.6: Additional Information and Full Hypothesis Test Examples - *CC BY 4.0*
  - 10.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet) - *CC BY 4.0*
  - 10.E: Hypothesis Testing with One Sample (Exercises) - *CC BY 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*