ToussaintPsyc350

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.

This text was compiled on 03/18/2025

TABLE OF CONTENTS

8: Tests of Means

9: Describing Bivariate Data

10: Regression

11: Power

12: Effect Size

13: Analysis of Variance

14: Case Studies

15: Calculators

Index

Glossary

Licensing

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

CHAPTER OVERVIEW

1: Introduction to Statistics

This first chapter begins by discussing what statistics are and why the study of statistics is important. Subsequent sections cover a variety of topics all basic to the study of statistics. One theme common to all of these sections is that they cover concepts and ideas important for other chapters in the book.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

1.1: What are Statistics?

Learning Objectives

- Describe the range of applications of statistics
- Identify situations in which statistics can be misleading
- Define "Statistics"

Statistics include numerical facts and figures. For instance:

- The largest earthquake measured 9.2 on the Richter scale.
- Men are at least 10 times more likely than women to commit murder.
- One in every 8 South Africans is HIV positive.
- By the year 2020, there will be 15 people aged 65 and over for every new baby born.

The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted. For example, consider the following three scenarios and the interpretations based upon the presented statistics. You will find that the numbers may be right, but the interpretation may be wrong. Try to identify a major flaw with each interpretation before we describe it.

1) A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.

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

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

A major flaw is that both increased churches and increased crime rates can be explained by larger populations. In bigger cities, there are both more churches and more crime. This problem, which we discuss in more detail in the section on Causation in Chapter 6, refers to the third-variable problem. Namely, a third variable can cause both situations; however, people erroneously believe that there is a causal relationship between the two primary variables rather than recognize that a third variable can cause both.

3) 75% more interracial marriages are occurring this year than 25 years ago. Thus, our society accepts interracial marriages.

A major flaw is that we don't have the information that we need. What is the rate at which marriages are occurring? Suppose only 1% of marriages 25 years ago were interracial and so now 1.75% of marriages are interracial (1.75 is 75% higher than 1). But this latter number is hardly evidence suggesting the acceptability of interracial marriages. In addition, the statistic provided does not rule out the possibility that the number of interracial marriages has seen dramatic fluctuations over the years and this year is not the highest. Again, there is simply not enough information to understand fully the impact of the statistics.

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

- Mikki Hebl

1.2: Importance of Statistics

Learning Objectives

- Give examples of statistics encountered in everyday life
- Give examples of how statistics can lend credibility to an argument

Like most people, you probably feel that it is important to "take control of your life." But what does this mean? Partly, it means being able to properly evaluate the data and claims that bombard you every day. If you cannot distinguish good from faulty reasoning, then you are vulnerable to manipulation and to decisions that are not in your best interest. Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, statistics is one of the most important things that you can study.

To be more specific, here are some claims that we have heard on several occasions. (We are not saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentine.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.
- Native Americans are significantly more likely to be hit crossing the street than are people of other ethnicities.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase one's life span.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over 400.
- There is an 80% chance that in a room full of 30 people that at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

All of these claims are statistical in character. We suspect that some of them sound familiar; if not, we bet that you have heard other claims like them. Notice how diverse the examples are. They come from psychology, health, law, sports, business, etc. Indeed, data and data interpretation show up in discourse from virtually every facet of contemporary life.

Statistics are often presented in an effort to add credibility to an argument or advice. You can see this by paying attention to television advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step towards taking control of your life. (It is not, of course, the only step needed for this purpose.) The present textbook is designed to help you learn statistical essentials. It will make you into an intelligent consumer of statistical claims.

You can take the first step right away. To be an intelligent consumer of statistics, your first reflex must be to question the statistics that you encounter. The British Prime Minister Benjamin Disraeli is quoted by Mark Twain as having said, "There are three kinds of lies -- lies, damned lies, and statistics." This quote reminds us why it is so important to understand statistics. So let us invite you to reform your statistical habits from now on. No longer will you blindly accept numbers or findings. Instead, you will begin to think about the numbers, their sources, and most importantly, the procedures used to generate them.

We have put the emphasis on defending ourselves against fraudulent claims wrapped up as statistics. We close this section on a more positive note. Just as important as detecting the deceptive use of statistics is the appreciation of the proper use of statistics. You must also learn to recognize statistical evidence that supports a stated conclusion. Statistics are all around you, sometimes used well, sometimes not. We must learn how to distinguish the two cases.

- Mikki Hebl

1.3: Descriptive Statistics

Learning Objectives

- Define "descriptive statistics"
- Distinguish between descriptive statistics and inferential statistics

Descriptive statistics are numbers that are used to summarize and describe data. The word "data" refers to the information that has been collected from an experiment, a survey, a historical record, etc. (By the way, "data" is plural. One piece of information is called a "datum.") If we are analyzing birth certificates, for example, a descriptive statistic might be the percentage of certificates issued in New York State, or the average age of the mother. Any other number we choose to compute also counts as a descriptive statistic for the data from which the statistic is computed. Several descriptive statistics are often used at one time, to give a full picture of the data.

Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand. Generalizing from our data to another set of cases is the business of inferential statistics, which you'll be studying in another Section. Here we focus on (mere) descriptive statistics. Some descriptive statistics are shown in Table 1.3.1. The table shows the average salaries for various occupations in the United States in 1999. (Click here to see how much individuals with other occupations earn.)

Table 1.3.1: Average salaries for various occupations in 1999.

Salary	Occupation
$112,760	pediatricians
$106,130	dentists
$100,090	podiatrists
$ 76,140	physicists
$ 53,410	architects
$ 49,720	school, clinical, and counseling psychologists
$ 47,910	flight attendants
$ 39,560	elementary school teachers
$ 38,710	police officers
$ 18,980	floral designers

Descriptive statistics like these offer insight into American society. It is interesting to note, for example, that we pay the people who educate our children and who protect our citizens a great deal less than we pay people who take care of our feet or our teeth.

For more descriptive statistics, consider Table 1.3.2 which shows the number of unmarried men per 100 unmarried women in U.S. Metro Areas in 1990. From this table we see that men outnumber women most in Jacksonville, NC, and women outnumber men most in Sarasota, FL. You can see that descriptive statistics can be useful if we are looking for an opposite-sex partner! (These data come from the Information Please Almanac.)

Table 1.3.2: Number of unmarried men per 100 *unmarried women in U.S. Metro Areas in* 1990.

Cities with mostly men	Men per 100 Women	Cities with mostly women	Men per 100 Women
1. Jacksonville, NC	224	1. Sarasota, FL	66
2. Killeen-Temple, TX	123	2. Bradenton, FL	68
3. Fayetteville, NC	118	3. Altoona, PA	69
4. Brazoria, TX	117	4. Springfield, IL	70
5. Lawton, OK	116	5. Jacksonville, TN	70

Cities with mostly men	Men per 100 Women	Cities with mostly women	Men per 100 Women
6. State College, PA	113	6. Gadsden, AL	70
7. Clarksville-Hopkinsville, TN-KY	113	7. Wheeling, WV	70
8. Anchorage, Alaska	112	8. Charleston, WV	71
9. Salinas-Seaside-Monterey, CA	112	9. St. Joseph, MO	71
10. Bryan-College Station, TX	111	10. Lynchburg, VA	71

NOTE: Unmarried includes never-married, widowed, and divorced persons, 15 years or older.

These descriptive statistics may make us ponder why the numbers are so disparate in these cities. One potential explanation, for instance, as to why there are more women in Florida than men may involve the fact that elderly individuals tend to move down to the Sarasota region and that women tend to outlive men. Thus, more women might live in Sarasota than men. However, in the absence of proper data, this is only speculation.

You probably know that descriptive statistics are central to the world of sports. Every sporting event produces numerous statistics such as the shooting percentage of players on a basketball team. For the Olympic marathon (a foot race of 26.2 miles), we possess data that cover more than a century of competition. (The first modern Olympics took place in 1896.) Table 1.3.3 shows the winning times for both men and women (the latter have only been allowed to compete since 1984).

Table 1.3.3: Winning Olympic marathon times.

Women			
Year	Winner	Country	Time
1984	Joan Benoit	USA	2:24:52
1988	Rosa Mota	POR	2:25:40
1992	Valentina Yegorova	UT	2:32:41
1996	Fatuma Roba	ETH	2:26:05
2000	Naoko Takahashi	JPN	2:23:14
2004	Mizuki Noguchi	JPN	2:26:20
Men			
Year	Winner	Country	Time
1896	Spiridon Louis	GRE	2:58:50
1900	Michel Theato	FRA	2:59:45
1904	Thomas Hicks	USA	3:28:53
1906	Billy Sherring	CAN	2:51:23
1908	Johnny Hayes	USA	2:55:18
1912	Kenneth McArthur	S. Afr.	2:36:54
1920	Hannes Kolehmainen	FIN	2:32:35
1924	Albin Stenroos	FIN	2:41:22
1928	Boughra El Ouafi	FRA	2:32:57
1932	Juan Carlos Zabala	ARG	2:31:36

Women			
1936	Sohn Kee-Chung	JPN	2:29:19
1948	Delfo Cabrera	ARG	2:34:51
1952	Emil Ztopek	CZE	2:23:03
1956	Alain Mimoun	FRA	2:25:00
1960	Abebe Bikila	ETH	2:15:16
1964	Abebe Bikila	ETH	2:12:11
1968	Mamo Wolde	ETH	2:20:26
1972	Frank Shorter	USA	2:12:19
1976	Waldemar Cierpinski	E.Ger	2:09:55
1980	Waldemar Cierpinski	E.Ger	2:11:03
1984	Carlos Lopes	POR	2:09:21
1988	Gelindo Bordin	ITA	2:10:32
1992	Hwang Young-Cho	S. Kor	2:13:23
1996	Josia Thugwane	S. Afr.	2:12:36
2000	Gezahenge Abera	ETH	2:10.10
2004	Stefano Baldini	ITA	2:10:55

There are many descriptive statistics that we can compute from the data in the table. To gain insight into the improvement in speed over the years, let us divide the men's times into two pieces, namely, the first 13 races (up to 1952) and the second 13 (starting from 1956). The mean winning time for the first 13 races is 2 hours, 44 minutes, and 22 seconds (written $2:44:22$). The mean winning time for the second 13 races is $2:13:18$. This is quite a difference (over half an hour). Does this prove that the fastest men are running faster? Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year? We can't answer this question with descriptive statistics alone. All we can affirm is that the two means are "suggestive."

Examining Table 3 leads to many other questions. We note that Takahashi (the lead female runner in 2000) would have beaten the male runner in 1956 and all male runners in the first 12 marathons. This fact leads us to ask whether the gender gap will close or remain constant. When we look at the times within each gender, we also wonder how much they will decrease (if at all) in the next century of the Olympics. Might we one day witness a sub-2 hour marathon? The study of statistics can help you make reasonable guesses about the answers to these questions.

- Mikki Hebl

1.4: Inferential Statistics

Learning Objectives

- Distinguish between a sample and a population
- Define inferential statistics
- Identify biased samples
- Distinguish between simple random sampling and stratified sampling
- Distinguish between random sampling and random assignment

Populations and samples

In statistics, we often rely on a sample, that is, a small subset of a larger set of data, to draw inferences about the larger set. The larger set is known as the population from which the sample is drawn.

Example 1.4.1

You have been hired by the National Election Commission to examine how the American people feel about the fairness of the voting procedures in the U.S. Whom will you ask?

It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans, and draw inferences about the entire country from their responses. The Americans actually queried constitute our sample of the larger population of all Americans. The mathematical procedures whereby we convert information about the sample into intelligent guesses about the population fall under the rubric of inferential statistics.

A sample is typically a small subset of the population. In the case of voting attitudes, we would sample a few thousand Americans drawn from the hundreds of millions that make up the country. In choosing a sample, it is therefore crucial that it not over-represent one kind of citizen at the expense of others. For example, something would be wrong with our sample if it happened to be made up entirely of Florida residents. If the sample held only Floridians, it could not be used to infer the attitudes of other Americans. The same problem would arise if the sample were comprised only of Republicans. Inferential statistics are based on the assumption that sampling is random. We trust a random sample to represent different segments of society in close to the appropriate proportions (provided the sample is large enough; see below).

Example 1.4.2

We are interested in examining how many math classes have been taken on average by current graduating seniors at American colleges and universities during their four years in school. Whereas our population in the last example included all US citizens, now it involves just the graduating seniors throughout the country. This is still a large set since there are thousands of colleges and universities, each enrolling many students. (New York University, for example, enrolls $48,000$ students.) It would be prohibitively costly to examine the transcript of every college senior. We therefore take a sample of college seniors and then make inferences to the entire population based on what we find. To make the sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our sample were 3.2. Then we might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors, or chose too many technical institutions that have heavy math requirements. Such bad sampling makes our sample unrepresentative of the population of all seniors.

To solidify your understanding of sampling bias, consider the following example. Try to identify the population and the sample, and then reflect on whether the sample is likely to yield the information desired.

Example 1.4.3

A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

In Example 1.4.3, the population consists of all students in the class. The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population.

Example 1.4.4

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

In Example 1.4.4, the population is the class of all freshmen at the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman; people who cannot do cartwheels probably did not volunteer! In the example, we are also not told of the gender of the volunteers. Were they all women, for example? That might affect the outcome, contributing to the non-representative nature of the sample (if the school is co-ed). Sampling Bias is Discussed in More Detail Here

Simple Random Sampling

Researchers adopt a variety of sampling strategies. The most straightforward is simple random sampling. Such sampling requires every member of the population to have an equal chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance. To check your understanding of simple random sampling, consider the following example. What is the population? What is the sample? Was the sample picked by simple random sampling? Is it biased?

Example 1.4.5

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

In Example 1.4.5, the population consists of all twins recorded in the National Twin Registry. It is important that the researcher only make statistical generalizations to the twins on this list, not to all twins in the nation or world. That is, the National Twin Registry may not be representative of all twins. Even if inferences are limited to the Registry, a number of problems affect the sampling procedure we described. For instance, choosing only twins whose last names begin with Z does not give every individual an equal chance of being selected into the sample. Moreover, such a procedure risks over-representing ethnic groups with many surnames that begin with Z. There are other reasons why choosing just the $Z's$ may bias the sample. Perhaps such people are more patient than average because they often find themselves at the end of the line! The same problem occurs with choosing twins whose last name begins with B. An additional problem for the $B's$ is that the “every-other-one” procedure disallowed adjacent names on the B part of the list from being both selected. Just this defect alone means the sample was not formed through simple random sampling.

Sample size matters

Recall that the definition of a random sample is a sample in which every member of the population has an equal chance of being selected. This means that the sampling procedure rather than the results of the procedure define what it means for a sample to be random. Random samples, especially if the sample size is small, are not necessarily representative of the entire population. For

example, if a random sample of 20 subjects were taken from a population with an equal number of males and females, there would be a nontrivial probability (0.06) that 70% or more of the sample would be female. (To see how to obtain this probability, see the section on the binomial distribution.) Such a sample would not be representative, although it would be drawn randomly. Only a large sample size makes it likely that our sample is close to representative of the population. For this reason, inferential statistics take into account the sample size when generalizing results from samples to populations. In later chapters, you'll see what kinds of mathematical techniques ensure this sensitivity to sample size.

More complex sampling

Sometimes it is not feasible to build a sample using simple random sampling. To see the problem, consider the fact that both Dallas and Houston are competing to be hosts of the 2012 Olympics. Imagine that you are hired to assess whether most Texans prefer Houston to Dallas as the host, or the reverse. Given the impracticality of obtaining the opinion of every single Texan, you must construct a sample of the Texas population. But now notice how difficult it would be to proceed by simple random sampling. For example, how will you contact those individuals who don't vote and don't have a phone? Even among people you find in the telephone book, how can you identify those who have just relocated to California (and had no reason to inform you of their move)? What do you do about the fact that since the beginning of the study, an additional 4,212 people took up residence in the state of Texas? As you can see, it is sometimes very difficult to develop a truly random procedure. For this reason, other kinds of sampling techniques have been devised. We now discuss two of them.

Random Assignment

In experimental research, populations are often hypothetical. For example, in an experiment comparing the effectiveness of a new anti-depressant drug with a placebo, there is no actual population of individuals taking the drug. In this case, a specified population of people with some degree of depression is defined and a random sample is taken from this population. The sample is then randomly divided into two groups; one group is assigned to the treatment condition (drug) and the other group is assigned to the control condition (placebo). This random division of the sample into two groups is called random assignment. Random assignment is critical for the validity of an experiment. For example, consider the bias that could be introduced if the first 20 subjects to show up at the experiment were assigned to the experimental group and the second 20 subjects were assigned to the control group. It is possible that subjects who show up late tend to be more depressed than those who show up early, thus making the experimental group less depressed than the control group even before the treatment was administered.

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

Stratified Sampling

Since simple random sampling often does not ensure a representative sample, a sampling method called stratified random sampling is sometimes used to make the sample more representative of the population. This method can be used if the population has a number of distinct "strata" or groups. In stratified sampling, you first identify members of your sample who belong to each group. Then you randomly sample from each of those subgroups in such a way that the sizes of the subgroups in the sample are proportional to their sizes in the population.

Let's take an example: Suppose you were interested in views of capital punishment at an urban university. You have the time and resources to interview 200 students. The student body is diverse with respect to age; many older people work during the day and enroll in night courses (average age is 39), while younger students generally enroll in day classes (average age of 19). It is possible that night students have different views about capital punishment than day students. If 70% of the students were day students, it makes sense to ensure that 70% of the sample consisted of day students. Thus, your sample of 200 students would consist of 140 day students and 60 night students. The proportion of day students in the sample and in the population (the entire university) would be the same. Inferences to the entire population of students at the university would therefore be more secure.

- Mikki Hebl and David Lane

1.5: Sampling Demonstration

Learning Objectives

- Distinguish between simple random sampling and stratified sampling.
- Describe how often random and stratified sampling give exactly the same result.

Review of Sampling

Instructions

The sampling simulation uses a population of 100 animals: 60 lions, 30 turtles, 10 rabbits.

Options

Random Sample: This option allows you to draw a sample of 10 animals at a time with each animal having an equal chance of being selected.

Stratified Sample: This option allows you to draw a sample of 10 animals at a time, with each number of animals from a group being proportional to their group's size of the population.

Simulation Results

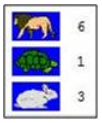


Figure 1.5.1: Simulation Results

The number of animals chosen from each group when a sample is drawn is shown next to the picture of the animal.

When you give it a try

Random Sampling

- Begin by leaving the Random Sample option selected.
- Click on the Select 10 Animals button, 10 animals will be selected out of the population.

  Note: The animals become highlighted in blue and a number count of each animal selected will be listed by each animal image.
- Each time you push the Select 10 Animals button, another sample will be drawn and the new tally will be shown on the right side of the previous sample.
- You should get different tally results for each animal as you select Select 10 Animals, however the computer may give you the same number drawn from an animal category every now and then.

Stratified Sample

Note: Your animals should become highlighted in blue and a number count should be listed by each animal image.

- Click on the Reset button, to clear the simulation.
- Select the Stratified Sampleoption.
- Click on the Select 10 Animals button a few times.
- As you get a new tally for every Select 10 Animals button, notice that the number of animals stays the same, but the animals selected are not always the same animals.

Illustrated Instructions

The opening screen of the sampling simulation displays all 100 animals in the population. You can select between a random sample and a stratified sample directly below the population and then generate a sample of ten animals.

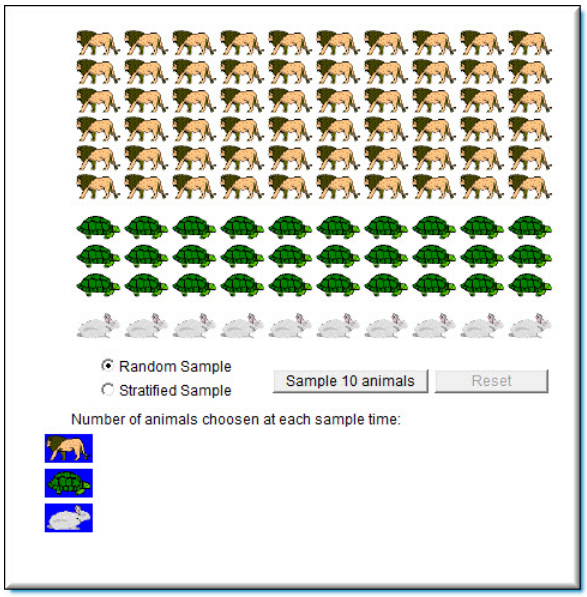


Figure 1.5.2: Sample Choices

Below is an example of a random sample. Notice that animals selected are highlighted in the population and the total number of animals selected from each category is listed at the bottom of the simulation.

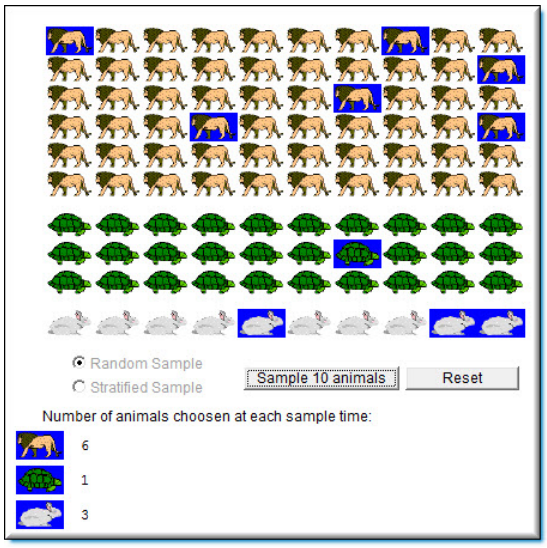


Figure 1.5.3: Random Sample

1.6: Variables

Learning Objectives

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

Independent and Dependent variables

Variables are properties or characteristics of some event, object, or person that can take on different values or amounts (as opposed to constants such as π that do not vary). When conducting research, experimenters often manipulate variables. For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is "type of antidepressant." When a variable is manipulated by an experimenter, it is called an independent variable. The experiment seeks to determine the effect of the *independent variable* on relief from depression. In this example, relief from depression is called a dependent variable. In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

Example 1.6.1

Can blueberries slow down aging? A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month-old rats (equivalent to 60-year-old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder. After eight weeks, the rats were given memory and motor skills tests. Although all supplemented rats showed improvement, those supplemented with blueberry powder showed the most notable improvement.

a. What is the independent variable?
b. What are the dependent variables?

Solution

a. dietary supplement: none, blueberry, strawberry, and spinach
b. memory test and motor skills test

More information on the blueberry study

Example 1.6.2

Does beta-carotene protect against cancer? Beta-carotene supplements have been thought to protect against cancer. However, a study published in the *Journal of the National Cancer Institute* suggests this is false. The study was conducted with $39,000$ women aged 45 and up. These women were randomly assigned to receive a beta-carotene supplement or a placebo, and their health was studied over their lifetime. Cancer rates for women taking the beta-carotene supplement did not differ systematically from the cancer rates of those women taking the placebo.

a. What is the independent variable?
b. What is the dependent variable?

Solution

a. supplements: beta-carotene or placebo
b. occurrence of cancer

Example 1.6.3

How bright is right? An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of a following car to realize that the car in front is stopping and to hit the brakes.

a. What is the independent variable?
b. What is the dependent variable?

Solution

a. brightness of brake lights
b. time to hit brakes

Levels of an Independent Variable: Experiments and Controls

If an experiment compares an experimental treatment with a control treatment, then the independent variable (type of treatment) has two levels: experimental and control. If an experiment were comparing five types of diets, then the independent variable (type of diet) would have five levels. In general, the number of levels of an independent variable is the number of experimental conditions.

Qualitative and Quantitative Variables

An important distinction between variables is between qualitative variables and quantitative variables. Qualitative variables are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. The values of a qualitative variable do not imply a numerical ordering. Values of the variable “religion” differ qualitatively; no ordering of religions is implied. Qualitative variables are sometimes referred to as categorical variables. Quantitative variables are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

In the study on the effect of diet discussed in Example $1.6.1$, the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The variable "type of supplement" is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable "memory test" is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

Discrete and Continuous Variables

Variables such as number of children in a household are called discrete variables since the possible scores are discrete points on the scale. For example, a household could have three children or six children, but not 4.53 children. Other variables such as "time to respond to a question" are continuous variables since the scale is continuous and not made up of discrete steps. The response time could be 1.64 seconds, or it could be 1.64237123922121seconds. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- Heidi Ziemer

1.7: Percentiles

Learning Objectives

- Define percentiles
- Use three formulas for computing percentiles

A test score in and of itself is usually difficult to interpret. For example, if you learned that your score on a measure of shyness was 35 out of a possible 50, you would have little idea how shy you are compared to other people. More relevant is the percentage of people with lower shyness scores than yours. This percentage is called a percentile. If 65% of the scores were below yours, then your score would be the 65th percentile.

Two Simple Definitions of Percentile

There is no universally accepted definition of a percentile. Using the 65th percentile as an example, the 65th percentile can be defined as the lowest score that is greater than 65% of the scores. This is the way we defined it above and we will call this " Definition 1." The 65th percentile can also be defined as the smallest score that is greater than or equal to 65% of the scores. This we will call "Definition 2." Unfortunately, these two definitions can lead to dramatically different results, especially when there is relatively little data. Moreover, neither of these definitions is explicit about how to handle rounding. For instance, what rank is required to be higher than 65% of the scores when the total number of scores is 50? This is tricky because 65% of 50 is 32.5. How do we find the lowest number that is higher than 32.5 of the scores? A third way to compute percentiles (presented below) is a weighted average of the percentiles computed according to the first two definitions. This third definition handles rounding more gracefully than the other two and has the advantage that it allows the median to be defined conveniently as the 50th percentile.

Third Definition

Unless otherwise specified, when we refer to "percentile," we will be referring to this third definition of percentiles. Let's begin with an example. Consider the 25th percentile for the 8 numbers in Table $1.7.1$. Notice the numbers are given ranks ranging from 1 for the lowest number to 8 for the highest number.

Table $1.7.1$: Test Scores.

Number	3	5	7	8	9	11	13	15
Rank	1	2	3	4	5	6	7	8

The first step is to compute the rank (R) of the 25th percentile. This is done using the following formula:

$$R = P/100 \times (N+1) \tag{1.7.1}$$

where P is the desired percentile (25 in this case) and N is the number of numbers (8 in this case). Therefore,

$$R = 25/100 \times (8+1) = 9/4 = 2.25 \tag{1.7.2}$$

If R is an integer, the Pth percentile is the number with rank R. When R is not an integer, we compute the Pth percentile by interpolation as follows:

1. Define IR as the integer portion of R (the number to the left of the decimal point). For this example, $IR = 2$.
2. Define FR as the fractional portion of R. For this example, $FR = 0.25$.
3. Find the scores with Rank IR and with Rank $IR+1$. For this example, this means the score with Rank 2 and the score with Rank 3. The scores are 5 and 7.
4. Interpolate by multiplying the difference between the scores by FR and add the result to the lower score. For these data, this is $(0.25)(7-5)+5 = 5.5$.

Therefore, the 25th percentile is 5.5. If we had used the first definition (the smallest score greater than 25% of the scores), the 25th percentile would have been 7. If we had used the second definition (the smallest score greater than or equal to 25% of the scores), the 25th percentile would have been 5.

For a second example, consider the 20 quiz scores shown in Table $1.7.2$.

Table 1.7.2: 20 *Quiz Scores.*

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

We will compute the 25th and the 85th percentiles. For the 25th,

$$R = 25/100 \times (20+1) = 21/4 = 5.25 \tag{1.7.3}$$

$$IR = 5 \ and \ FR = 0.25 \tag{1.7.4}$$

Since the score with a rank of IR (which is 5) and the score with a rank of $IR+1$ (which is 6) are both equal to 5, the 25th percentile is 5. In terms of the formula:

$$25^{th} \text{ percentile} = (0.25) \times (5-5) + 5 = 5 \tag{1.7.5}$$

For the 85th percentile,

$$R = 85/100 \times (20+1) = 17.85. \tag{1.7.6}$$

$$IR = 17 \ and \ FR = 0.85 \tag{1.7.7}$$

Caution: FR does not generally equal the percentile to be computed as it does here.

The score with a rank of 17 is 9 and the score with a rank of 18 is 10. Therefore, the 85th percentile is:

$$(0.85)(10-9) + 9 = 9.85 \tag{1.7.8}$$

Consider the 50th percentile of the numbers $2, 3, 5, 9$.

$$R = 50/100 \times (4+1) = 2.5 \tag{1.7.9}$$

$$IR = 2 \ and \ FR = 0.5 \tag{1.7.10}$$

The score with a rank of IR is 3 and the score with a rank of $IR+1$ is 5. Therefore, the 50th percentile is:

$$(0.5)(5-3) + 3 = 4 \tag{1.7.11}$$

Finally, consider the 50th percentile of the numbers $2, 3, 5, 9, 11$.

$$R = 50/100 \times (5+1) = 3 \tag{1.7.12}$$

$$IR = 3 \ and \ FR = 0 \tag{1.7.13}$$

Whenever $FR = 0$, you simply find the number with rank IR. In this case, the third number is equal to 5, so the 50th percentile is 5. You will also get the right answer if you apply the general formula:

$$50^{th} \text{ percentile} = (0.00)(9-5) + 5 = 5 \tag{1.7.14}$$

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- David M. Lane

1.8: Levels of Measurement

Learning Objectives

- Define and distinguish among nominal, ordinal, interval, and ratio scales
- Identify a scale type
- Discuss the type of scale used in psychological measurement
- Give examples of errors that can be made by failing to understand the proper use of measurement scales

Types of Scales

Before we can conduct a statistical analysis, we need to measure our dependent variable. Exactly how the measurement is carried out depends on the type of variable involved in the analysis. Different types are measured differently. To measure the time taken to respond to a stimulus, you might use a stop watch. Stop watches are of no use, of course, when it comes to measuring someone's attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like "very favorable," "somewhat favorable," etc.). For a dependent variable such as "favorite color," you can simply note the color-word (like "red") that the subject offers.

Although procedures for measurement differ in many ways, they can be classified using a few fundamental categories. In a given category, all of the procedures share some properties that are important for you to know about. The categories are called "scale types," or just "scales," and are described in this section.

Nominal scales

When measuring using a nominal scale, one simply names or categorizes responses. Gender, handedness, favorite color, and religion are examples of variables measured on a nominal scale. The essential point about nominal scales is that they do not imply any ordering among the responses. For example, when classifying people according to their favorite color, there is no sense in which green is placed "ahead of" blue. Responses are merely categorized. Nominal scales embody the lowest level of measurement.

Ordinal scales

A researcher wishing to measure consumers' satisfaction with their microwave ovens might ask them to specify their feelings as either "very dissatisfied," "somewhat dissatisfied," "somewhat satisfied," or "very satisfied." The items in this scale are ordered, ranging from least to most satisfied. This is what distinguishes ordinal from nominal scales. Unlike nominal scales, ordinal scales allow comparisons of the degree to which two subjects possess the dependent variable. For example, our satisfaction ordering makes it meaningful to assert that one person is more satisfied than another with their microwave ovens. Such an assertion reflects the first person's use of a verbal label that comes later in the list than the label chosen by the second person.

On the other hand, ordinal scales fail to capture important information that will be present in the other scales we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels. In our satisfaction scale, for example, the difference between the responses "very dissatisfied" and "somewhat dissatisfied" is probably not equivalent to the difference between "somewhat dissatisfied" and "somewhat satisfied." Nothing in our measurement procedure allows us to determine whether the two differences reflect the same difference in psychological satisfaction. Statisticians express this point by saying that the differences between adjacent scale values do not necessarily represent equal intervals on the underlying scale giving rise to the measurements. (In our case, the underlying scale is the true feeling of satisfaction, which we are trying to measure.)

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

Interval scales

Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).

Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name "zero." The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label "zero" is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures. For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the two ratios. After all, if the "zero" label were applied at the temperature that Fahrenheit happens to label as 10 degrees, the two ratios would instead be 30 to 10 and 90 to 40, no longer the same! For this reason, it does not make sense to say that 80 degrees is "twice as hot" as 40 degrees. Such a claim would depend on an arbitrary decision about where to "start" the temperature scale, namely, what temperature to call zero (whereas the claim is intended to make a more fundamental assertion about the underlying physical reality).

Ratio scales

The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured. You can think of a ratio scale as the three earlier scales rolled up in one. Like a nominal scale, it provides a name or category for each object (the numbers serve as labels). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference at two places on the scale has the same meaning. And in addition, the same ratio at two places on the scale also carries the same meaning.

The Fahrenheit scale for temperature has an arbitrary zero point and is therefore not a ratio scale. However, zero on the Kelvin scale is absolute zero. This makes the Kelvin scale a ratio scale. For example, if one temperature is twice as high as another as measured on the Kelvin scale, then it has twice the kinetic energy of the other temperature.

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

What level of measurement is used for psychological variables?

Rating scales are used frequently in psychological research. For example, experimental subjects may be asked to rate their level of pain, how much they like a consumer product, their attitudes about capital punishment, their confidence in an answer to a test question. Typically these ratings are made on a 5-point or a 7-point scale. These scales are ordinal scales since there is no assurance that a given difference represents the same thing across the range of the scale. For example, there is no way to be sure that a treatment that reduces pain from a rated pain level of 3 to a rated pain level of 2 represents the same level of relief as a treatment that reduces pain from a rated pain level of 7 to a rated pain level of 6.

In memory experiments, the dependent variable is often the number of items correctly recalled. What scale of measurement is this? You could reasonably argue that it is a ratio scale. First, there is a true zero point: some subjects may get no items correct at all. Moreover, a difference of one represents a difference of one item recalled across the entire scale. It is certainly valid to say that someone who recalled 12 items recalled twice as many items as someone who recalled only 6 items.

But number-of-items recalled is a more complicated case than it appears at first. Consider the following example in which subjects are asked to remember as many items as possible from a list of 10. Assume that (a) there are 5 easy items and 5 difficult items, (b) half of the subjects are able to recall all the easy items and different numbers of difficult items, while (c) the other half of the subjects are unable to recall any of the difficult items but they do remember different numbers of easy items. Some sample data are shown below.

Table 1.8.1

Subject	Easy Items	Difficult Items	**Score**

Subject	Easy Items					Difficult Items					**Score**
A	0	0	1	1	0	0	0	0	0	0	**2**
B	1	0	1	1	0	0	0	0	0	0	**3**
C	1	1	1	1	1	1	1	0	0	0	**7**
D	1	1	1	1	1	0	1	1	0	1	**8**

Let's compare (1) the difference between Subject $A's$ score of 2 and Subject $B's$ score of 3 with (2) the difference between Subject $C's$ score of 7 and Subject $D's$ score of 8. The former difference is a difference of one easy item; the latter difference is a difference of one difficult item. Do these two differences necessarily signify the same difference in memory? We are inclined to respond "No" to this question since only a little more memory may be needed to retain the additional easy item whereas a lot more memory may be needed to retain the additional hard item. The general point is that it is often inappropriate to consider psychological measurement scales as either interval or ratio.

Consequences of level of measurement

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

Table $1.8.2$

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

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

Table $1.8.3$

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

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

Does it make sense to compute the mean of numbers measured on an ordinal scale? This is a difficult question, one that statisticians have debated for decades. You will be able to explore this issue yourself in a simulation shown in the next section and reach your own conclusion. The prevailing (but by no means unanimous) opinion of statisticians is that for almost all practical situations, the mean of an ordinally-measured variable is a meaningful statistic. However, as you will see in the simulation, there are extreme situations in which computing the mean of an ordinally-measured variable can be very misleading.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- Dan Osherson and David M. Lane

1.9: Measurements

Learning Objectives

- Understand what it means for a scale to be ordinal and its relationship to interval scales.
- Determine whether an investigator can be misled by computing the means of an ordinal scale.

Instructions

This is a demonstration of a very complex issue. Experts in the field disagree on how to interpret differences on an ordinal scale, so do not be discouraged if it takes you a while to catch on. In this demonstration you will explore the relationship between interval and ordinal scales. The demonstration is based on two brands of baked goods.

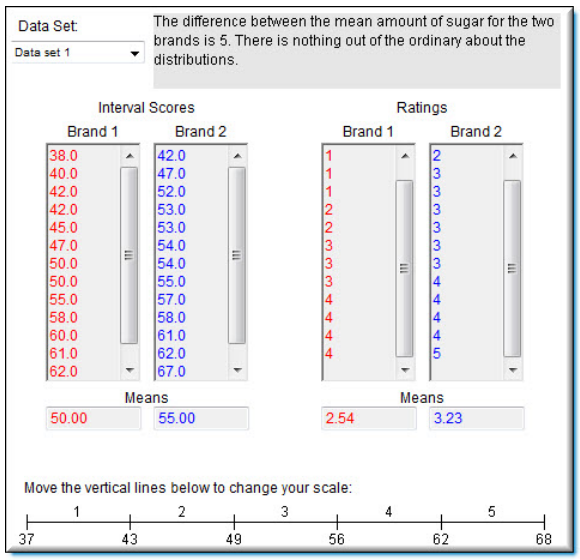


Figure 1.9.1: Interval scores and Ratings for two brands

The data on the left side labeled "interval scores" shows the amount of sugar in each of 12 products. The column labeled " Brand 1" contains the sugar content of each of 12 brand-one products. The second column ("Brand 1") shows the sugar content of the brand-two products. The amount of sugar is measured on an interval scale.

A rater tastes each of the products and rates them on a 5-point "sweetness" scale. Rating scales are typically ordinal rather than interval.

The scale at the bottom shows the "mapping" of sugar content onto the ratings. Sugar content between 37 and 43 is rated as 1, between 43 and $49, 2, etc$. Therefore, the difference between a rating of 1 and a rating of 2 represents, on average a "sugar difference" of 6. A difference between a rating of 2 and a rating of 3 also represents, on average a "sugar difference" of 6. The original ratings are rounded off and displayed are on an interval scale. It is likely that rater's ratings would not be on an interval scale. You can change the cutoff points between ratings by moving the vertical lines with the mouse. As you change these cutoffs, the ratings change automatically. For example, you might see what the ratings would look like if people did not consider something very sweet (rating of 5) unless it was very very sweet.

The mean amount of sugar in Data Set 1 is 50 for the first brand and 55 for the second brand. The obvious conclusion is that, on average, the second brand is sweeter than the first. However, pretend that you only had the ratings to go by and were not aware of the actual amounts of sugar. Would you reach the correct decision if you compared the mean ratings of the two brands. Change the cutoffs for mapping the interval sugar scale onto the ordinal rating scale. Do any mappings lead to incorrect interpretations? Try this with Data Set 1 and with Data Set 2. Try to find a situation where the mean sweetness rating is higher for Brand 2 even though the mean amount of sugar is greater for Brand 1. If you find such a situation, then you have found an instance in which using the means of ordinal data lead to incorrect conclusions. It is possible to find this situation, so look hard.

Keep in mind that in realistic situations, you only know the ratings and not the "true" interval scale that underlies them. If you knew the interval scale, you would use it.

1.10: Distributions

Learning Objectives

- Define "distribution"
- Interpret a frequency distribution
- Distinguish between a frequency distribution and a probability distribution
- Construct a grouped frequency distribution for a continuous variable
- Identify the skew of a distribution
- Identify bimodal, leptokurtic, and platykurtic distributions

Distributions of Discrete Variables

A recently purchased a bag of Plain M&M's contained candies of six different colors. A quick count showed that there were 55 M&M's: 17 brown, 18 red, 7 yellow, 7 green, 2 blue, and 4 orange. These counts are shown below in Table $1.10.1$.

Table 1.10.1: Frequencies in the Bag of M&M's

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

This table is called a frequency table and it describes the distribution of M&M color frequencies. Not surprisingly, this kind of distribution is called a frequency distribution. Often a frequency distribution is shown graphically as in Figure $1.10.1$.

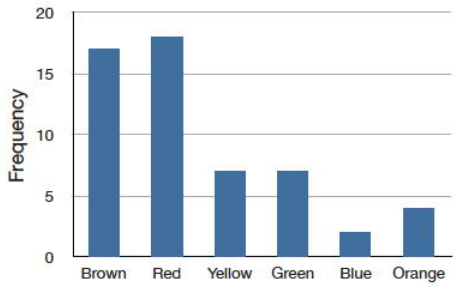


***Figure* $1.10.1$:** *Distribution of* 55 *M&M's.*

The distribution shown in Figure $1.10.1$ concerns just my one bag of M&M's. You might be wondering about the distribution of colors for all M&M's. The manufacturer of M&M's provides some information about this matter, but they do not tell us exactly how many M&M's of each color they have ever produced. Instead, they report proportions rather than frequencies. Figure $1.10.2$ shows these proportions. Since every M&M is one of the six familiar colors, the six proportions shown in the figure add to one. We call Figure $1.10.2$ a probability distribution because if you choose an M&M at random, the probability of getting, say, a brown M&M is equal to the proportion of M&M's that are brown (0.30).

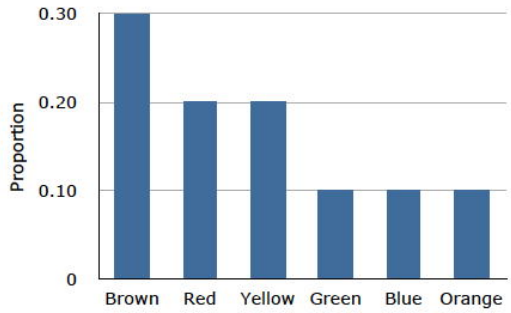


Figure 1.10.2: Distribution of all M&M's.

Notice that the distributions in Figures 1.10.1 and 1.10.2 are not identical. Figure 1.10.1 portrays the distribution in a sample of 55 M&M's. Figure 1.10.2 shows the proportions for all M&M's. Chance factors involving the machines used by the manufacturer introduce random variation into the different bags produced. Some bags will have a distribution of colors that is close to Figure 1.10.2; others will be further away.

Continuous Variables

The variable "color of M&M" used in this example is a discrete variable, and its distribution is also called discrete. Let us now extend the concept of a distribution to continuous variables. The data shown in Table 1.10.2 are the times it took one of us (DL) to move the mouse over a small target in a series of 20 trials. The times are sorted from shortest to longest. The variable "time to respond" is a continuous variable. With time measured accurately (to many decimal places), no two response times would be expected to be the same. Measuring time in milliseconds (thousandths of a second) is often precise enough to approximate a continuous variable in Psychology. As you can see in Table 1.10.2, measuring DL's responses this way produced times no two of which were the same. As a result, a frequency distribution would be uninformative: it would consist of the 20 times in the experiment, each with a frequency of 1.

Table 1.10.2: Response Times

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

The solution to this problem is to create a grouped frequency distribution. In a grouped frequency distribution, scores falling within various ranges are tabulated. Table 1.10.3 shows a grouped frequency distribution for these 20 times.

Table 1.10.3: Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5

Range	Frequency
900-1000	0
1000-1100	1

Grouped frequency distributions can be portrayed graphically. Figure $1.10.3$ shows a graphical representation of the frequency distribution in Table $1.10.3$. This kind of graph is called a histogram. A later chapter contains an entire section devoted to histograms.

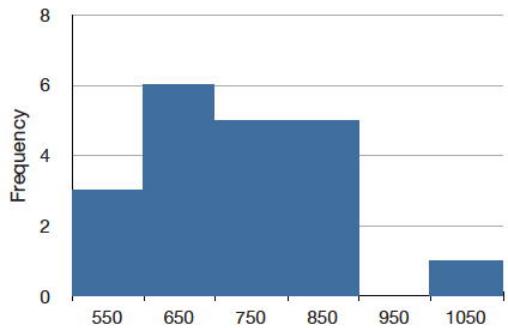


Figure 1.10.3: A histogram of the grouped frequency distribution shown in Table 1.10.3. The labels on the X*-axis are the middle values of the range they represent.*

Probability Densities

The histogram in Figure $1.10.3$ portrays just DL's 20 times in the one experiment he performed. To represent the probability associated with an arbitrary movement (which can take any positive amount of time), we must represent all these potential times at once. For this purpose, we plot the distribution for the continuous variable of time. Distributions for continuous variables are called continuous distributions. They also carry the fancier name probability density. Some probability densities have particular importance in statistics. A very important one is shaped like a bell, and called the normal distribution. Many naturally-occurring phenomena can be approximated surprisingly well by this distribution. It will serve to illustrate some features of all continuous distributions.

An example of a normal distribution is shown in Figure $1.10.4$. Do you see the "bell"? The normal distribution doesn't represent a real bell, however, since the left and right tips extend indefinitely (we can't draw them any further so they look like they've stopped in our diagram). The Y-axis in the normal distribution represents the "density of probability." Intuitively, it shows the chance of obtaining values near corresponding points on the X-axis. In Figure $1.10.4$, for example, the probability of an observation with value near 40 is about half of the probability of an observation with value near 50. (For more information, please see the chapter on normal distributions.)

Although this text does not discuss the concept of probability density in detail, you should keep the following ideas in mind about the curve that describes a continuous distribution (like the normal distribution). First, the area under the curve equals 1. Second, the probability of any exact value of X is 0. Finally, the area under the curve and bounded between two given points on the X-axis is the probability that a number chosen at random will fall between the two points. Let us illustrate with DL's hand movements. First, the probability that his movement takes some amount of time is one! (We exclude the possibility of him never finishing his gesture.) Second, the probability that his movement takes exactly 598.956432342346576 milliseconds is essentially zero. (We can make the probability as close as we like to zero by making the time measurement more and more precise.) Finally, suppose that the probability of DL's movement taking between 600 and 700 milliseconds is one tenth. Then the continuous distribution for DL's possible times would have a shape that places 10% of the area below the curve in the region bounded by 600 and 700 on the X-axis.

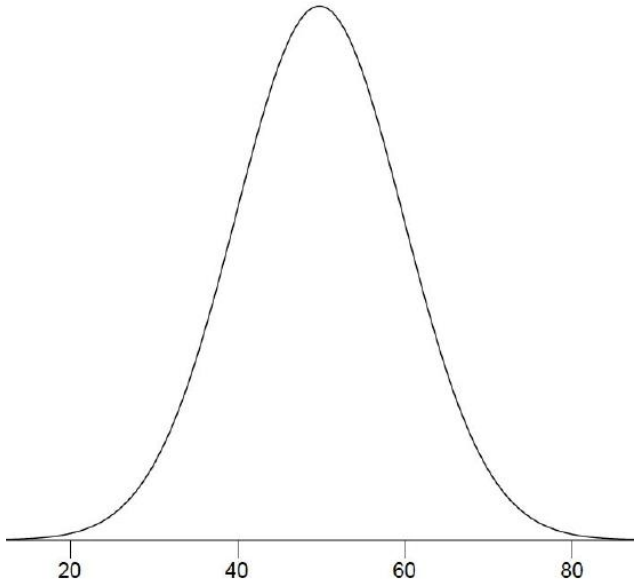


Figure 1.10.4: A normal distribution.

Shapes of Distributions

Distributions have different shapes; they don't all look like the normal distribution in Figure 1.10.4. For example, the normal probability density is higher in the middle compared to its two tails. Other distributions need not have this feature. There is even variation among the distributions that we call "normal." For example, some normal distributions are more spread out than the one shown in Figure 1.10.4 (their tails begin to hit the X-axis further from the middle of the curve -- for example, at 10 and 90 if drawn in place of Figure 1.10.4). Others are less spread out (their tails might approach the X-axis at 30 and 70). More information on the normal distribution can be found in a later chapter completely devoted to them.

The distribution shown in Figure 1.10.4 is symmetric; if you folded it in the middle, the two sides would match perfectly. Figure 1.10.5 shows the discrete distribution of scores on a psychology test. This distribution is not symmetric: the tail in the positive direction extends further than the tail in the negative direction. A distribution with the longer tail extending in the positive direction is said to have a positive skew. It is also described as "skewed to the right."

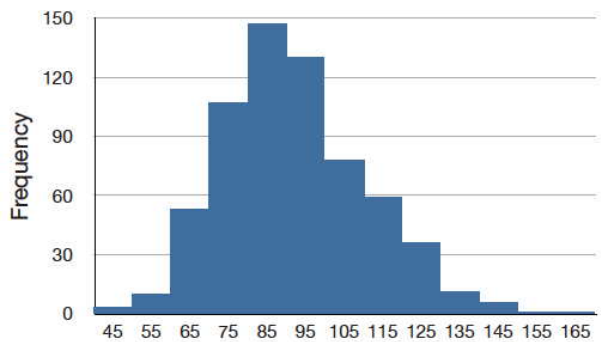


Figure 1.10.5: (left) A distribution with a positive skew.

Figure 1.10.6 shows the salaries of major league baseball players in 1974 (in thousands of dollars). This distribution has an extreme positive skew.

Figure 1.10.6: (right) A distribution with a very large positive skew.

A continuous distribution with a positive skew is shown in Figure 1.10.7.

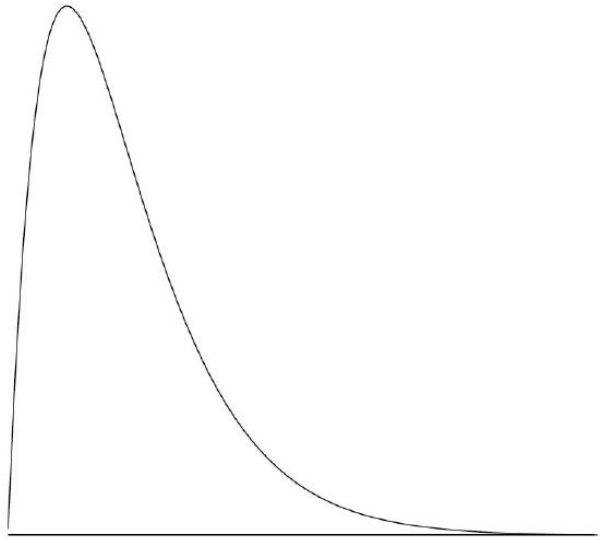


Figure 1.10.7: A continuous distribution with a positive skew.

Although less common, some distributions have a negative skew. Figure 1.10.8 shows the scores on a 20-point problem on a statistics exam. Since the tail of the distribution extends to the left, this distribution is skewed to the left.

Figure 1.10.8: A distribution with negative skew.

The histogram in Figure 1.10.8 shows the frequencies of various scores on a 20-point question on a statistics test.

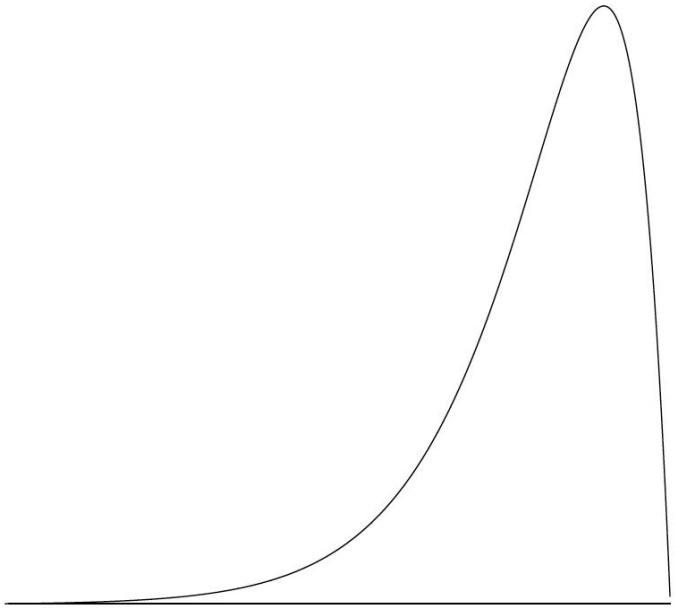


Figure 1.10.9: A continuous distribution with a negative skew.

A continuous distribution with a negative skew is shown in Figure 1.10.9. The distributions shown so far all have one distinct high point or peak. The distribution in Figure 1.10.10 has two distinct peaks. A distribution with two peaks is called a bimodal distribution.

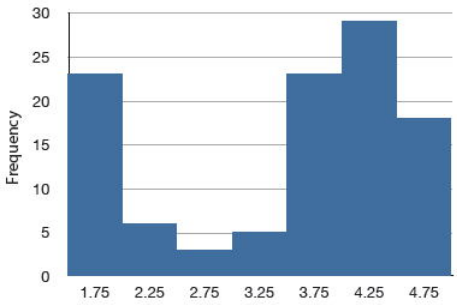


Figure 1.10.10. Frequencies of times between eruptions of the Old Faithful geyser. Notice the two distinct peaks: one at 1.75 and the other at 4.25.

Distributions also differ from each other in terms of how large or "fat" their tails are. Figure 1.10.11 shows two distributions that differ in this respect. The upper distribution has relatively more scores in its tails; its shape is called leptokurtic. The lower distribution has relatively fewer scores in its tails; its shape is called platykurtic.

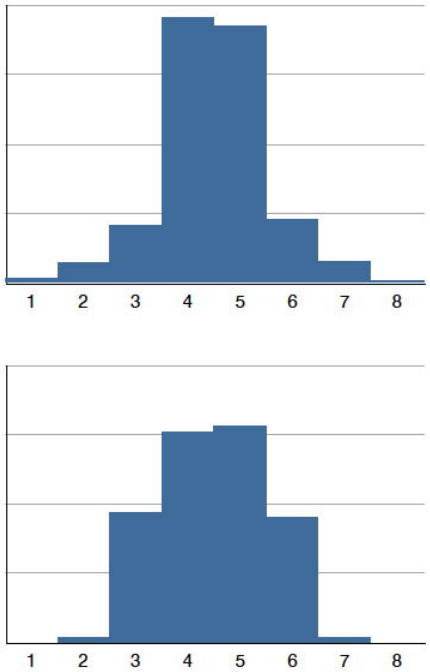


Figure 1.10.11. Distributions differing in kurtosis. The top distribution has long tails. It is called "leptokurtic." The bottom distribution has short tails. It is called "platykurtic."

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- David M. Lane and Heidi Ziemer

1.11: Summation Notation

Learning Objectives

- Use summation notation to express the sum of all numbers
- Use summation notation to express the sum of a subset of numbers
- Use summation notation to express the sum of squares

Many statistical formulas involve summing numbers. Fortunately there is a convenient notation for expressing summation. This section covers the basics of this summation notation.

Let's say we have a variable X that represents the weights (in grams) of 4 grapes. The data are shown in Table $1.11.1$.

Table 1.11.1: *Weights of* 4 *grapes.*

Grape	X
1	4.6
2	5.1
3	4.9
4	4.4

We label Grape $1's$ weight X_1, Grape $2's$ weight X_2, etc. The following formula means to sum up the weights of the four grapes:

$$\sum_{i=1}^{4} X_i \tag{1.11.1}$$

The Greek letter capital sigma ($\sum$) indicates summation. The "$i=1$" at the bottom indicates that the summation is to start with X_1 and the 4 at the top indicates that the summation will end with X_4. The "X_i" indicates that X is the variable to be summed as i goes from 1 to 4. Therefore,

$$\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4 = 4.6 + 5.1 + 4.9 + 4.4 = 19.0 \tag{1.11.2}$$

The symbol

$$\sum_{i=1}^{3} X_i \tag{1.11.3}$$

indicates that only the first 3 scores are to be summed. The index variable i goes from 1 to 3.

When all the scores of a variable (such as X) are to be summed, it is often convenient to use the following abbreviated notation:

$$\sum X \tag{1.11.4}$$

Thus, when no values of i are shown, it means to sum all the values of X.

Many formulas involve squaring numbers before they are summed. This is indicated as

$$\sum X^2 = 4.62 + 5.12 + 4.92 + 4.42 = 21.16 + 26.01 + 24.01 + 19.36 = 90.54 \tag{1.11.5}$$

Notice that:

$$\left(\sum X\right)^2 \neq \sum X^2 \tag{1.11.6}$$

because the expression on the left means to sum up all the values of X and then square the sum ($19^2 = 361$), whereas the expression on the right means to square the numbers and then sum the squares (90.54, as shown).

Some formulas involve the sum of cross products. Table $1.11.2$ shows the data for variables X and Y. The cross products (XY) are shown in the third column. The sum of the cross products is $3+4+21=28$.

Table 1.11.2*: Cross Products.*

X	Y	XY
1	3	3
2	2	4
3	7	21

In summation notation, this is written as:

$$\sum XY = 28. \tag{1.11.7}$$

- David M. Lane

1.12: Linear Transformations

Learning Objectives

- Give the formula for a linear transformation
- Determine whether a transformation is linear
- Describe what is linear about a linear transformation

Often it is necessary to transform data from one measurement scale to another. For example, you might want to convert height measured in feet to height measured in inches. Table $1.12.1$ shows the heights of four people measured in both feet and inches. To transform feet to inches, you simply multiply by 12. Similarly, to transform inches to feet, you divide by 12.

Table 1.12.1: Converting between feet and inches

Feet	Inches
5.00	60
6.25	75
5.50	66
5.75	69

Some conversions require that you multiply by a number and then add a second number. A good example of this is the transformation between degrees Celsius and degrees Fahrenheit. Table $1.12.2$ shows the temperatures of five US cities in the early afternoon of November $16, 2002$.

Table 1.12.2: Temperatures in 5 cities on $11/16/2002$

City	Degrees Fahrenheit	Degrees Celsius
Houston	54	12.22
Chicago	37	2.78
Minneapolis	31	-0.56
Miami	78	25.56
Phoenix	70	21.11

The formula to transform Celsius to Fahrenheit is:

$$F = 1.8C + 32 \tag{1.12.1}$$

The formula for converting from Fahrenheit to Celsius is

$$C = 0.5556F - 17.778 \tag{1.12.2}$$

The transformation consists of multiplying by a constant and then adding a second constant. For the conversion from Celsius to Fahrenheit, the first constant is 1.8 and the second is 32.

Figure $1.12.1$ shows a plot of degrees Celsius as a function of degrees Fahrenheit. Notice that the points form a straight line. This will always be the case if the transformation from one scale to another consists of multiplying by one constant and then adding a second constant. Such transformations are therefore called linear transformations.

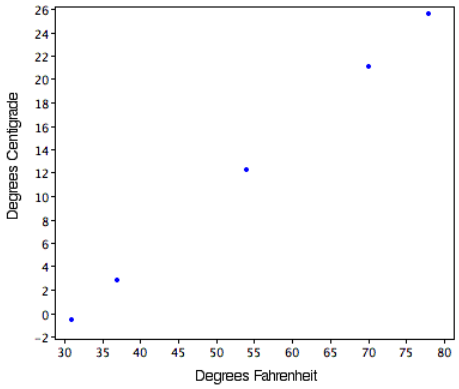


Figure 1.12.1*: Degrees Celsius as a function of degrees Fahrenheit*

Many transformations are not linear. With nonlinear transformations, the points in a plot of the transformed variable against the original variable would not fall on a straight line. Examples of nonlinear transformations are: square root, raising to a power, logarithm, and any of the trigonometric functions.

- David M. Lane

1.13: Logarithms

Learning Objectives

- Compute logs using different bases
- Perform basic arithmetic operations using logs
- State the relationship between logs and proportional change

The log transformation reduces positive skew. This can be valuable both for making the data more interpretable and for helping to meet the assumptions of inferential statistics.

Basics of Logarithms (Logs)

Logs are, in a sense, the opposite of exponents. Consider the following simple expression:

$$10^2 = 100 \tag{1.13.1}$$

Here we can say the base of 10 is raised to the second power. Here is an example of a log:

$$\log_{10}(100) = 2 \tag{1.13.2}$$

This can be read as: The log base ten of 100 equals 2. The result is the power that the base of 10 has to be raised to in order to equal the value (100). Similarly,

$$\log_{10}(1000) = 3 \tag{1.13.3}$$

since 10 has to be raised to the third power in order to equal $1,000$.

These examples all used base 10, but any base could have been used. There is a base which results in "natural logarithms" and that is called e and equals approximately 2.718. It is beyond the scope here to explain what is "natural" about it. Natural logarithms can be indicated either as: $\ln(x)\ or\ \log_e(x)$.

Changing the base of the log changes the result by a multiplicative constant. To convert from $\log_{10}$ to natural logs, you multiply by 2.303. Analogously, to convert in the other direction, you divide by 2.303.

$$\ln X = 2.303 \log_{10} X \tag{1.13.4}$$

Taking the antilog of a number undoes the operation of taking the $\log$. Therefore, since $\log_{10}(1000) = 3$, the $antilog_{10}$ of 3 is $10^3 = 1,000$. Taking the antilog of a number simply raises the base of the logarithm in question to that number.

Logs and Proportional Change

A series of numbers that increase proportionally will increase in equal amounts when converted to logs. For example, the numbers in the first column of Table $1.13.1$
increase by a factor of 1.5 so that each row is 1.5 times as high as the preceding row. The $\log_{10}$ transformed numbers increase in equal steps of 0.176.

Table 1.13.1: *Proportional raw changes are equal in log units*

Raw	Log
4.0	0.602
6.0	0.778
9.0	0.954
13.5	1.130

As another example, if one student increased their score from 100 to 200 while a second student increased theirs from 150 to 300, the percentage change (100%) is the same for both students. The log difference is also the same, as shown below.

$$
\begin{aligned}
Log_{10}(100) &= 2.000 \\
\log_{10}(200) &= 2.301 \\
Difference &: 0.301
\end{aligned}
\tag{1.13.5}
$$

$$
\begin{aligned}
\log_{10}(150) &= 2.176 \\
\log_{10}(300) &= 2.477 \\
Difference &: 0.301
\end{aligned}
$$

Arithmetic Operations

Rules for logs of products and quotients are shown below.

$$\log(AB) = \log(A) + \log(B) \tag{1.13.6}$$

$$\log\left(\frac{A}{B}\right) = \log(A) - \log(B) \tag{1.13.7}$$

For example,

$$\log_{10}(10 \times 100) = \log_{10}(10) + \log_{10}(100) = 1 + 2 = 3. \tag{1.13.8}$$

Similarly,

$$\log_{10}\left(\frac{100}{10}\right) = \log_{10}(100) - \log_{10}(10) = 2 - 1 = 1. \tag{1.13.9}$$

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- David M. Lane

1.14: Statistical Literacy

Do Athletes Get Special Treatment?

Prerequisites

Levels of Measurement

Figure 1.14.1: Runners

The Board of Trustees at a university commissioned a top management-consulting firm to address the admission processes for academic and athletic programs. The consulting firm wrote a report discussing the trade-off between maintaining academic and athletic excellence. One of their key findings was:

> The standard for an athlete's admission, as reflected in SAT scores alone, is lower than the standard for non-athletes by as much as 20 percent, with the weight of this difference being carried by the so-called "revenue sports" of football and basketball. Athletes are also admitted through a different process than the one used to admit non-athlete students.

What do you think?

Based on what you have learned in this chapter about measurement scales, does it make sense to compare SAT scores using percentages? Why or why not?

> As you may know, the SAT has an arbitrarily-determined lower limit on test scores of 200. Therefore, SAT is measured on either an ordinal scale or, at most, an interval scale. However, it is clearly not measured on a ratio scale. Therefore, it is not meaningful to report SAT score differences in terms of percentages. For example, consider the effect of subtracting 200 from every student's score so that the lowest possible score is 0. How would that affect the difference as expressed in percentages?

Statistical Errors in Politics

Prerequisites

Inferential Statistics

Figure 1.14.2*: Survey*

An article about ignorance of statistics in politics quotes a politician commenting on why the "American Community Survey" should be eliminated:

> "We're spending $70 per person to fill this out. That's just not cost effective, especially since in the end this is not a scientific survey. It's a random survey."

What do you think?

What is wrong with this statement? Despite the error in this statement, what type of sampling could be done so that the sample will be more likely to be representative of the population?

> Randomness is what makes the survey scientific. If the survey were not random, then it would be biased and therefore statistically meaningless, especially since the survey is conducted to make generalizations about the American population. Stratified sampling would likely be more representative of the population.

Reference

Mark C. C., scientopia.org

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- Denise Harvey and David Lane

1.E: Introduction to Statistics (Exercises)

General Questions

Q1

A teacher wishes to know whether the males in his/her class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared. Is this an example of descriptive or inferential statistics? (relevant section 1, relevant section 2)

Q2

A cognitive psychologist is interested in comparing two ways of presenting stimuli on subsequent memory. Twelve subjects are presented with each method and a memory test is given. What would be the roles of descriptive and inferential statistics in the analysis of these data? (relevant section 1 & relevant section 2)

Q3

If you are told that you scored in the 80th percentile, from just this information would you know exactly what that means and how it was calculated? Explain. (relevant section)

Q4

A study is conducted to determine whether people learn better with spaced or massed practice. Subjects volunteer from an introductory psychology class. At the beginning of the semester 12 subjects volunteer and are assigned to the massed-practice condition. At the end of the semester 12 subjects volunteer and are assigned to the spaced-practice condition. This experiment involves two kinds of non-random sampling:

1. Subjects are not randomly sampled from some specified population
2. Subjects are not randomly assigned to conditions.

Which of the problems relates to the generality of the results? Which of the problems relates to the validity of the results? Which problem is more serious? (relevant section)

Q5

Give an example of an independent and a dependent variable. (relevant section)

Q6

Categorize the following variables as being qualitative or quantitative: (relevant section)

a. Rating of the quality of a movie on a 7-point scale
b. Age
c. Country you were born in
d. Favorite Color
e. Time to respond to a question

Q7

Specify the level of measurement used for the items in Question 6. (relevant section)

Q8

Which of the following are linear transformations? (relevant section)

a. Converting from meters to kilometers
b. Squaring each side to find the area
c. Converting from ounces to pounds
d. Taking the square root of each person's height.
e. Multiplying all numbers by 2 and then adding 5
f. Converting temperature from Fahrenheit to Centigrade

Q9

The formula for finding each student's test grade (g) from his or her raw score (s) on a test is as follows:

$$g = 16 + 3s \tag{1.E.1}$$

Is this a linear transformation? If a student got a raw score of 20, what is his test grade? (relevant section)

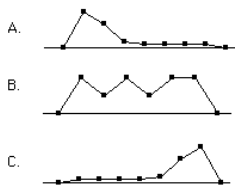
Q10

For the numbers $1, 2, 4, 16$, compute the following: (relevant section)

a. $\sum X$
b. $\sum X^2$
c. $\left(\sum X\right)^2$

Q11

Which of the frequency polygons has a large positive skew? Which has a large negative skew? (relevant section)



Q12

What is more likely to have a skewed distribution: time to solve an anagram problem (where the letters of a word or phrase are rearranged into another word or phrase like "dear" and "read" or "funeral" and "real fun") or scores on a vocabulary test? (relevant section)

Questions from Case Studies:

The following questions are from the Angry Moods (AM) case study.

Q13

(AM#1) Which variables are the participant variables? (They act as independent variables in this study.) (relevant section)

Q14

(AM#2) What are the dependent variables? (relevant section)

Q15

(AM#3) Is Anger-Out a quantitative or qualitative variable? (relevant section)

The following question is from the Teacher Ratings (TR) case study.

Q16

(TR#1) What is the independent variable in this study? (relevant section)

The following questions are from the ADHD Treatment (AT) case study.

Q17

(AT#1) What is the independent variable of this experiment? How many levels does it have? (relevant section)

Q18

(AT#2) What is the dependent variable? On what scale (nominal, ordinal, interval, ratio) was it measured? (relevant section)

Select Answers

S9

76

S10

$23,277,529$

CHAPTER OVERVIEW

2: Graphing Distributions

Graphing data is the first and often most important step in data analysis. In this day of computers, researchers all too often see only the results of complex computer analyses without ever taking a close look at the data themselves. This is all the more unfortunate because computers can create many types of graphs quickly and easily. This chapter covers some classic types of graphs such as bar charts that were invented by William Playfair in the 18th century as well as graphs such as box plots invented by John Tukey in the 20th century.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

2.1: Graphing Qualitative Variables

Learning Objectives

- Create a frequency table
- Determine when pie charts are valuable and when they are not
- Create and interpret bar charts
- Identify common graphical mistakes

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owner, a previous Windows owner, or a new computer purchaser.

This section examines graphical methods for displaying the results of the interviews. We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous Macintosh users. This situation may be contrasted with quantitative data, such as a person's weight. People of one weight are naturally ordered with respect to people of a different weight.

Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table $2.1.1$ shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for "none" is $85/500 = 0.17$.

Table* $2.1.1$*: *Frequency Table for the iMac Data*

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

Pie Charts

The pie chart in Figure $2.1.1$ shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

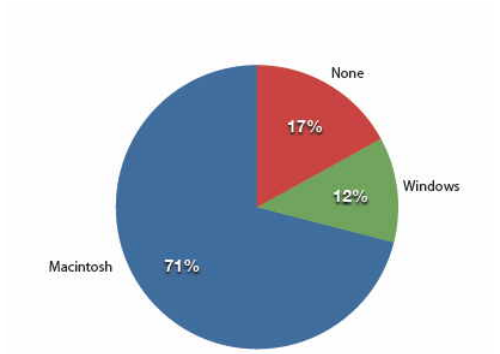


Figure 2.1.1: Pie chart of iMac purchases illustrating frequencies of previous computer ownership

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted, "The only worse design than a pie chart is several of them."

Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have occurred since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

Bar charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in Figure $2.1.2$. Frequencies are shown on the Y-axis and the type of computer previously owned is shown on the X-axis. Typically, the Y-axis shows the number of observations in each category rather than the percentage of observations as is typical in pie charts.

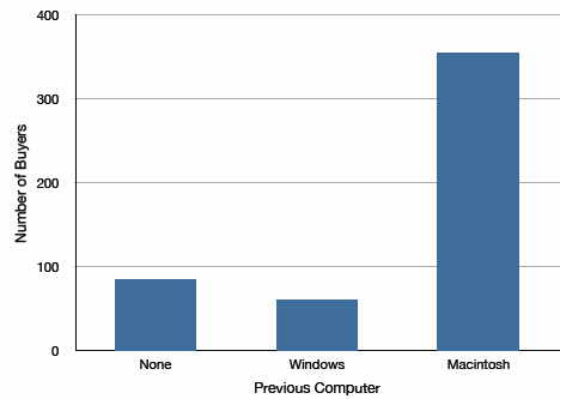


Figure $2.1.2$: Bar chart of iMac purchases as a function of previous computer ownership

Comparing Distributions

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the "distributions" of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure $2.1.3$ shows the number of people playing card games at the Yahoo website on a Sunday and on a Wednesday in the Spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

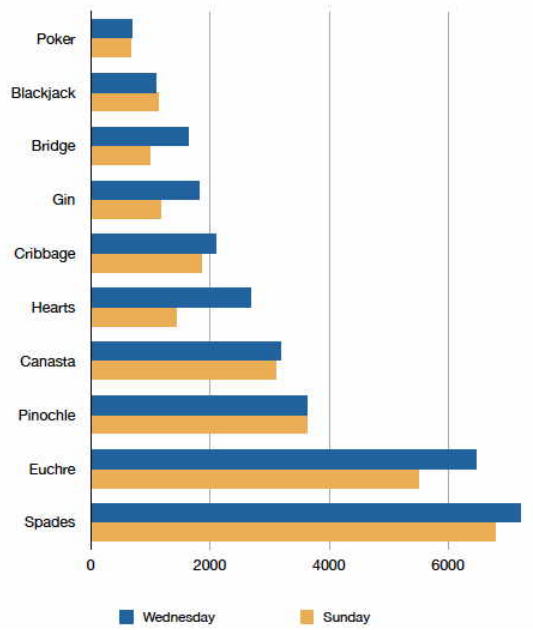


Figure 2.1.3: A bar chart of the number of people playing different card games on Sunday and Wednesday

The bars in Figure 2.1.3 are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. We'll have more to say about bar charts when we consider numerical quantities later in the section Bar Charts.

Some graphical mistakes to avoid

Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts such as the one shown in Figure 2.1.4 are usually not as effective as their two-dimensional counterparts.

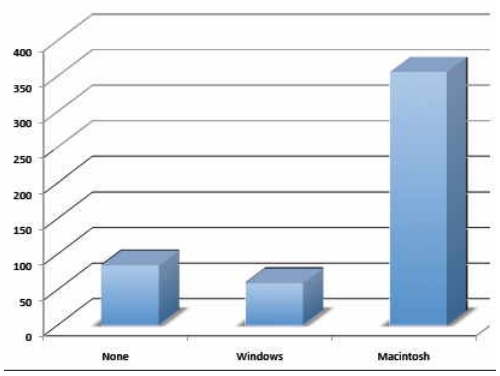


Figure 2.1.2

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure 2.1.5 presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 2.1.5 is misleading because the viewer's attention will be captured by areas. The areas can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in Figure 2.1.5 is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure 2.1.5 instead of Figure 2.1.2! Edward Tufte coined the term "lie factor" to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.

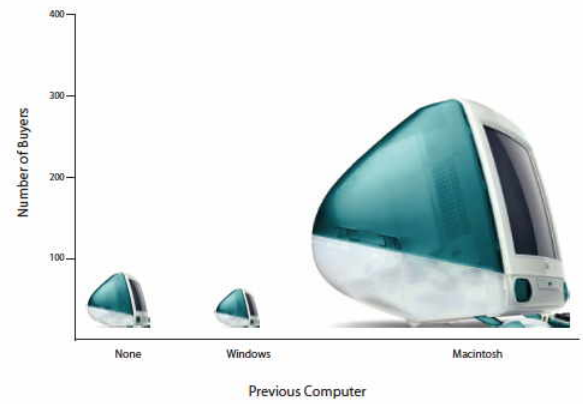


Figure 2.1.2 with a lie factor greater than 8

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, but not always, this number should be zero. Figure 2.1.6 shows the iMac data with a baseline of 50. Once again, the differences in areas suggest a different story than the true differences in percentages. The percentage of Windows-switchers seems minuscule compared to its true value of 12%.

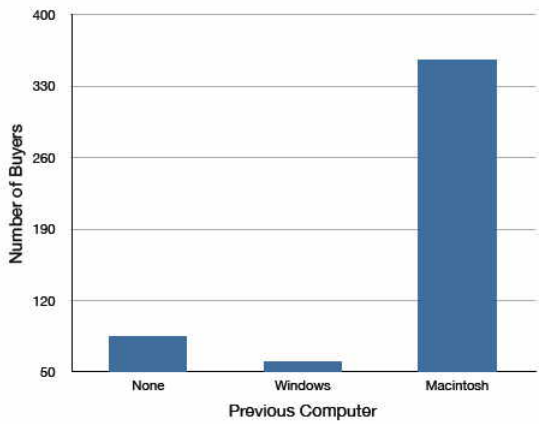


Figure 2.1.2 with a baseline of 50

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure 2.1.7 inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure 2.1.7 is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

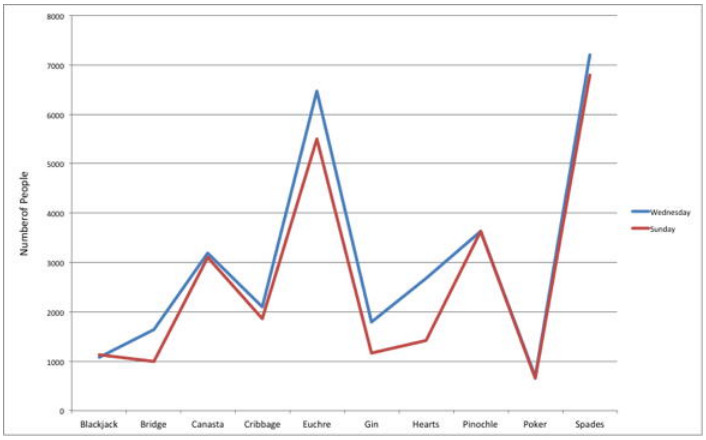


***Figure* 2.1.7:** *A line graph used inappropriately to depict the number of people playing different card games on Sunday and Wednesday.*

Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.

2.2: Quantitative Variables

As discussed in the section on variables in Chapter 1, quantitative variables are variables measured on a numeric scale. Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of quantitative variables. Quantitative variables are distinguished from categorical (sometimes called qualitative) variables such as favorite color, religion, city of birth, and favorite sport in which there is no ordering or measuring involved.

There are many types of graphs that can be used to portray distributions of quantitative variables. The upcoming sections cover the following types of graphs:

1. stem and leaf displays
2. histograms
3. frequency polygons
4. box plots
5. bar charts
6. line graphs
7. scatter plots (discussed in a different chapter)
8. dot plots

Some graph types such as stem and leaf displays are best-suited for small to moderate amounts of data, whereas others such as histograms are best-suited for large amounts of data. Graph types such as box plots are good at depicting differences between distributions. Scatter plots are used to show the relationship between two variables.

2.3: Stem and Leaf Displays

Learning Objectives

- Create and interpret basic stem and leaf displays
- Create and interpret back-to-back stem and leaf displays
- Judge whether a stem and leaf display is appropriate for a given data set

A stem and leaf display is a graphical method of displaying data. It is particularly useful when your data are not too numerous. In this section, we will explain how to construct and interpret this kind of graph.

As usual, an example will get us started. Consider Table $2.3.1$ that shows the number of touchdown passes (TD passes) thrown by each of the 31 teams in the National Football League in the 2000 season.

***Table* $2.3.1$:** *Number of touchdown passes*

$$\begin{array}{ccccccccc} 37 & 33 & 33 & 32 & 29 & 28 & 28 & 23 & 22 \\ 22 & 22 & 21 & 21 & 21 & 20 & 20 & 19 & 19 \\ 18 & 18 & 18 & 18 & 16 & 15 & 14 & 14 & 14 \\ 12 & 12 & 9 & 6 & & & & & \end{array} \tag{2.3.1}$$

A stem and leaf display of the data is shown in Figure $2.3.1$. The left portion of Figure $2.3.1$ contains the stems. They are the numbers $3, 2, 1,\ and\ 0$, arranged as a column to the left of the bars. Think of these numbers as $10's$ digits. A stem of 3, for example, can be used to represent the $10's$ digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leaves, and they represent the $1's$ digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

$$\begin{array}{c|cccccccccccc} 3 & 2 & 3 & 3 & 7 & & & & & & & & \\ 2 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 3 & 8 & 8 & 9 \\ 1 & 2 & 2 & 4 & 4 & 4 & 5 & 6 & 8 & 8 & 8 & 8 & 9 & 9 \\ 0 & 6 & 9 & & & & & & & & & & \end{array} \tag{2.3.2}$$

***Figure* $2.3.1$:** *Stem and leaf display of the number of touchdown passes*

To make this clear, let us examine Figure $2.3.1$ more closely. In the top row, the four leaves to the right of stem 3 are $2, 3, 3,\ and\ 7$. Combined with the stem, these leaves represent the numbers $32, 33, 33,\ and\ 37$, which are the numbers of TD passes for the first four teams in Table $2.3.1$. The next row has a stem of 2 and 12 leaves. Together, they represent 12 data points, namely, two occurrences of 20 TD passes, three occurrences of 21 TD passes, three occurrences of 22 TD passes, one occurrence of 23 TD passes, two occurrences of 28 TD passes, and one occurrence of 29 TD passes. We leave it to you to figure out what the third row represents. The fourth row has a stem of 0 and two leaves. It stands for the last two entries in Table $2.3.1$, namely 9 TD passes and 6 TD passes. (The latter two numbers may be thought of as 09 and 06.)

One purpose of a stem and leaf display is to clarify the shape of the distribution. You can see many facts about TD passes more easily in Figure $2.3.1$ than in Table $2.3.1$. For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TDs, with a few having more and a few having less. The precise numbers of TD passes can be determined by examining the leaves.

We can make our figure even more revealing by splitting each stem into two parts. Figure $2.3.2$ shows how to do this. The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in Table $2.3.1$. The second row is reserved for the numbers from 30 to 34 and holds the $32, 33,\ and\ 33$ TD passes made by the next three teams in the table. You can see for yourself what the other rows represent.

$$\begin{array}{c|cccccccccc} 3 & 7 & & & & & & & & \\ 3 & 2 & 3 & 3 & & & & & & \\ 2 & 8 & 8 & 9 & & & & & & \\ 2 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 3 \\ 1 & 5 & 6 & 8 & 8 & 8 & 8 & 9 & 9 & \\ 1 & 2 & 2 & 4 & 4 & 4 & & & & \\ 0 & 6 & 9 & & & & & & & \end{array} \tag{2.3.3}$$

***Figure* $2.3.2$:** *Stem and leaf display with the stems split in two*

Figure $2.3.2$ is more revealing than Figure $1PageIndex2$ because the latter figure lumps too many values into a single row. Whether you should split stems in a display depends on the exact form of your data. If rows get too long with single stems, you might try splitting them into two or more parts.

There is a variation of stem and leaf displays that is useful for comparing distributions. The two distributions are placed back to back along a common column of stems. The result is a "back-to-back stem and leaf graph." Figure $2.3.3$ shows such a graph. It compares the numbers of TD passes in the 1998 and 2000 seasons. The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data. For example, the second-to-last row shows that in 1998 there were teams with $11, 12,\ and\ 13$ TD passes, and in 2000 there were two teams with 12 and three teams with 14 TD passes.

$$\begin{array}{r|c|cccccccccc} 1\ 1 & 4 & & & & & & & & & \\ & 3 & 7 & & & & & & & & \\ 3\ 3\ 2 & 3 & 2 & 3 & 3 & & & & & & \\ 8\ 8\ 6\ 5 & 2 & 8 & 8 & 9 & & & & & & \\ 4\ 4\ 3\ 3\ 1\ 1\ 1\ 0 & 2 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 3 \\ 9\ 8\ 7\ 7\ 7\ 6\ 6\ 6\ 5 & 1 & 5 & 6 & 8 & 8 & 8 & 8 & 9 & 9 & \\ 3\ 2\ 1 & 1 & 2 & 2 & 4 & 4 & 4 & 4 & & & \\ 7 & 0 & 6 & 9 & & & & & & & \end{array} \tag{2.3.4}$$

***Figure* $2.3.3$:** *Back-to-back stem and leaf display.*

The left side shows the 1998 TD data and the right side shows the 2000 TD data. Figure $2.3.3$ helps us see that the two seasons were similar, but that only in 1998 did any teams throw more than 40 TD passes.

There are two things about the football data that make them easy to graph with stems and leaves. First, the data are limited to whole numbers that can be represented with a one-digit stem and a one-digit leaf. Second, all the numbers are positive. If the data include numbers with three or more digits, or contain decimals, they can be rounded to two-digit accuracy.

Negative values are also easily handled. Let us look at another example.

Table $2.3.2$ shows data from the case study Weapons and Aggression. Each value is the mean difference over a series of trials between the times it took an experimental subject to name aggressive words (like “punch”) under two conditions. In one condition, the words were preceded by a non-weapon word such as "bug." In the second condition, the same words were preceded by a weapon word such as "gun" or "knife." The issue addressed by the experiment was whether a preceding weapon word would speed up (or prime) pronunciation of the aggressive word compared to a non-weapon priming word. A positive difference implies greater priming of the aggressive word by the weapon word. Negative differences imply that the priming by the weapon word was less than for a neutral word.

***Table* $2.3.2$:** *The effects of priming (thousandths of a second)*

$$\begin{array}{cccccccc} 43.2 & 42.9 & 35.6 & 25.6 & 25.4 & 23.6 & & \\ 20.5 & 19.9 & 14.4 & 12.7 & 11.3 & 10.2 & & \\ 10.0 & 9.1 & 7.5 & 5.4 & 4.7 & 3.8 & 2.1 & 1.2 \\ -0.2 & -6.3 & -6.7 & -8.8 & -10.4 & -10.5 & & \\ -14.9 & -14.9 & -15.0 & -18.5 & -27.4 & & & \end{array} \tag{2.3.5}$$

You see that the numbers range from 43.2 to -27.4. The first value indicates that one subject was 43.2 milliseconds faster pronouncing aggressive words when they were preceded by weapon words than when preceded by neutral words. The value -27.4 indicates that another subject was 27.4 milliseconds slower pronouncing aggressive words when they were preceded by weapon words.

The data are displayed with stems and leaves in Figure $2.3.4$. Since stem and leaf displays can only portray two whole digits (one for the stem and one for the leaf), the numbers are first rounded. Thus, the value 43.2 is rounded to 43 and represented with a stem of 4 and a leaf of 3. Similarly, 42.9 is rounded to 43. To represent negative numbers, we simply use negative stems. For example, the bottom row of the figure represents the number -27. The second-to-last row represents the numbers $-10, -10, -15$, etc. Once again, we have rounded the original values from Table $2.3.2$.

$$\begin{array}{r|lllllll} 4 & 3 & 3 & & & & & \\ 3 & 6 & & & & & & \\ 2 & 0 & 0 & 4 & 5 & 6 & & \\ 1 & 0 & 0 & 1 & 3 & 4 & & \\ 0 & 1 & 2 & 4 & 5 & 5 & 8 & 9 \\ -0 & 0 & 6 & 7 & 9 & & & \\ -1 & 0 & 0 & 5 & 5 & 5 & 9 & \\ -2 & 7 & & & & & & \end{array} \tag{2.3.6}$$

***Figure* $2.3.4$:** *Stem and leaf display with negative numbers and rounding*

Observe that the figure contains a row headed by "0" and another headed by "-0". The stem of 0 is for numbers between 0 and 9, whereas the stem of -0 is for numbers between 0 and -9. For example, the fifth row of the table holds the numbers $1, 2, 4, 5, 5, 8, 9$and the sixth row holds $0, -6, -7, \ and \ -9$. Values that are exactly 0 before rounding should be split as evenly as possible between the "0" and "-0" rows. In Table $2.3.2$, none of the values are 0 before rounding. The "0" that appears in the "-0" row comes from the original value of -0.2 in the table.

Although stem and leaf displays are unwieldy for large data sets, they are often useful for data sets with up to 200 observations. Figure $2.3.5$ portrays the distribution of populations of 185 US cities in 1998. To be included, a city had to have between $100,000$and $500,000$residents.

$$\begin{array}{r|lllllllllllllllllllllllllllllllllllllllllllllllll} 4 & 8 & 9 & 9 \\ 4 & 6 \\ 4 & 4 & 4 & 5 & 5 \\ 4 & 3 & 3 & 3 \\ 4 & 0 & 1 \\ 3 & 9 & 9 \\ 3 & 6 & 7 & 7 & 7 & 7 & 7 \\ 3 & 5 & 5 \\ 3 & 2 & 2 & 3 \\ 3 & 1 & 1 & 1 \\ 2 & 8 & 8 & 9 & 9 \\ 2 & 6 & 6 & 6 & 6 & 6 & 7 \\ 2 & 4 & 4 & 4 & 4 & 5 & 5 \\ 2 & 2 & 2 & 3 & 3 & 3 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 \\ 1 & 6 & 6 & 6 & 6 & 6 & 6 & 7 & 7 & 7 & 7 & 7 & 7 \\ 1 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \tag{2.3.7}$$

***Figure* $2.3.5$:** *Stem and leaf display of populations of* 185 *US cities with populations between* $100,000$*and* $500,000$*in* 1998.

Since a stem and leaf plot shows only two-place accuracy, we had to round the numbers to the nearest $10,000$. For example, the largest number ($493,559$) was rounded to $490,000$and then plotted with a stem of 4 and a leaf of 9. The fourth highest number ($463,201$) was rounded to $460,000$and plotted with a stem of 4 and a leaf of 6. Thus, the stems represent units of $100,000$and the leaves represent units of $10,000$. Notice that each stem value is split into five parts: $0-1, 2-3, 4-5, 6-7$, and $8-9$.

Whether your data can be suitably represented by a stem and leaf graph depends on whether they can be rounded without loss of important information. Also, their extreme values must fit into two successive digits, as the data in Figure $2.3.5$ fit into the $10,000$ and $100,000$ places (for leaves and stems, respectively). Deciding what kind of graph is best suited to displaying your data thus requires good judgment. Statistics is not just recipes!

2.4: Histograms

Learning Objectives

- Create a grouped frequency distribution
- Create a histogram based on a grouped frequency distribution
- Determine an appropriate bin width

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psychology test. The test consists of 197 items, each graded as "correct" or "incorrect." The students' scores ranged from 46 to 167.

The first step is to create a frequency table. Unfortunately, a simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table $2.4.1$.

Table $2.4.1$: *Grouped Frequency Distribution of Psychology Test Scores*

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

To create this table, the range of scores was broken into intervals, called class intervals. The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval was counted to obtain the class frequencies. There are three scores in the first interval, 10 in the second, etc.

Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too "choppy." More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. A histogram of these data is shown in Figure $2.4.1$.

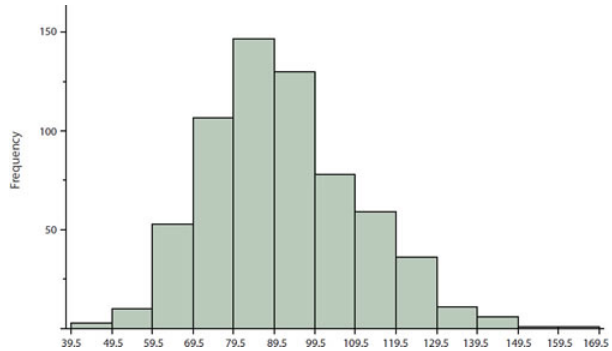


Figure 2.4.1: Histogram of scores on a psychology test.

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not symmetric: the scores extend to the right farther than they do to the left. The distribution is therefore said to be skewed. (We'll have more to say about shapes of distributions in the chapter " Summarizing Distributions.")

In our example, the observations are whole numbers. Histograms can also be used when the scores are measured on a more continuous scale such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence-sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, $4000,\ 5000$, etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance, and is the practice of many computer programs that create histograms. Note also that some computer programs label the middle of each interval rather than the end points.

Histograms can be based on relative frequencies instead of actual frequencies. Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. In this case, the Y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the Y-axis (labeled as proportion).

Sturges' rule

There is more to be said about the widths of the class intervals, sometimes called bin widths. Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. There are some "rules of thumb" that can help you choose an appropriate width. (But keep in mind that none of the rules is perfect.) Sturges' rule is to set the number of intervals as close as possible to $1+\log_2(N)$, where $\log_2(N)$ is the base 2 log of the number of observations. The formula can also be written as $1+3.3\log_2(N)$, where $\log_{10}(N)$ is the log base 10 of the number of observations. According to Sturges' rule, 1000 observations would be graphed with 11 class intervals since 10 is the closest integer to $\log_2(1000)$. We prefer the Rice rule, which is to set the number of intervals to twice the cube root of the number of observations. In the case of 1000 observations, the Rice rule yields 20 intervals instead of the 11 recommended by Sturges' rule. For the psychology test example used above, Sturges' rule recommends 10 intervals while the Rice rule recommends 17. In the end, we compromised and chose 13 intervals for Figure $2.4.1$ to create a histogram that seemed clearest. The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.

To provide experience in constructing histograms, we have developed an interactive demonstration. The demonstration reveals the consequences of different choices of bin width and of lower boundary for the first interval.

Interactive histogram

2.5: Frequency Polygons

Learning Objectives

- Create and interpret frequency polygons
- Create and interpret cumulative frequency polygons
- Create and interpret overlaid frequency polygons

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

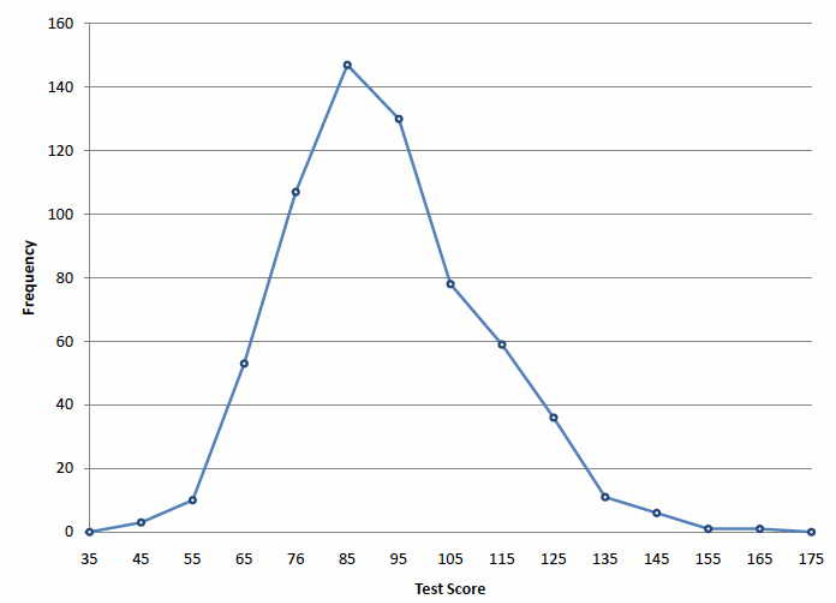


Figure 2.5.1: Frequency polygon for the psychology test scores

A frequency polygon for 642 psychology test scores shown in Figure 2.5.1 was constructed from the frequency table shown in Table 2.5.1.

Table 2.5.1: Frequency Distribution of Psychology Test Scores.

Lower Limit	Upper Limit	Count	Cumulative Count
29.5	39.5	0	0
39.5	49.5	3	3
49.5	59.5	10	13
59.5	69.5	53	66
69.5	79.5	107	173
79.5	89.5	147	320
89.5	99.5	130	450
99.5	109.5	78	528

Lower Limit	Upper Limit	Count	Cumulative Count
109.5	119.5	59	587
119.5	129.5	36	623
129.5	139.5	11	634
139.5	149.5	6	640
149.5	159.5	1	641
159.5	169.5	1	642
169.5	179.5	0	642

The first label on the X-axis is 35. This represents an interval extending from 29.5 to 39.5. Since the lowest test score is 46, this interval has a frequency of 0. The point labeled 45 represents the interval from 39.5 to 49.5. There are three scores in this interval. There are 147 scores in the interval that surrounds 85.

You can easily discern the shape of the distribution from Figure $2.5.1$. Most of the scores are between 65 and 115. It is clear that the distribution is not symmetric inasmuch as good scores (to the right) trail off more gradually than poor scores (to the left). In the terminology of Chapter 3 (where we will study shapes of distributions more systematically), the distribution is skewed.

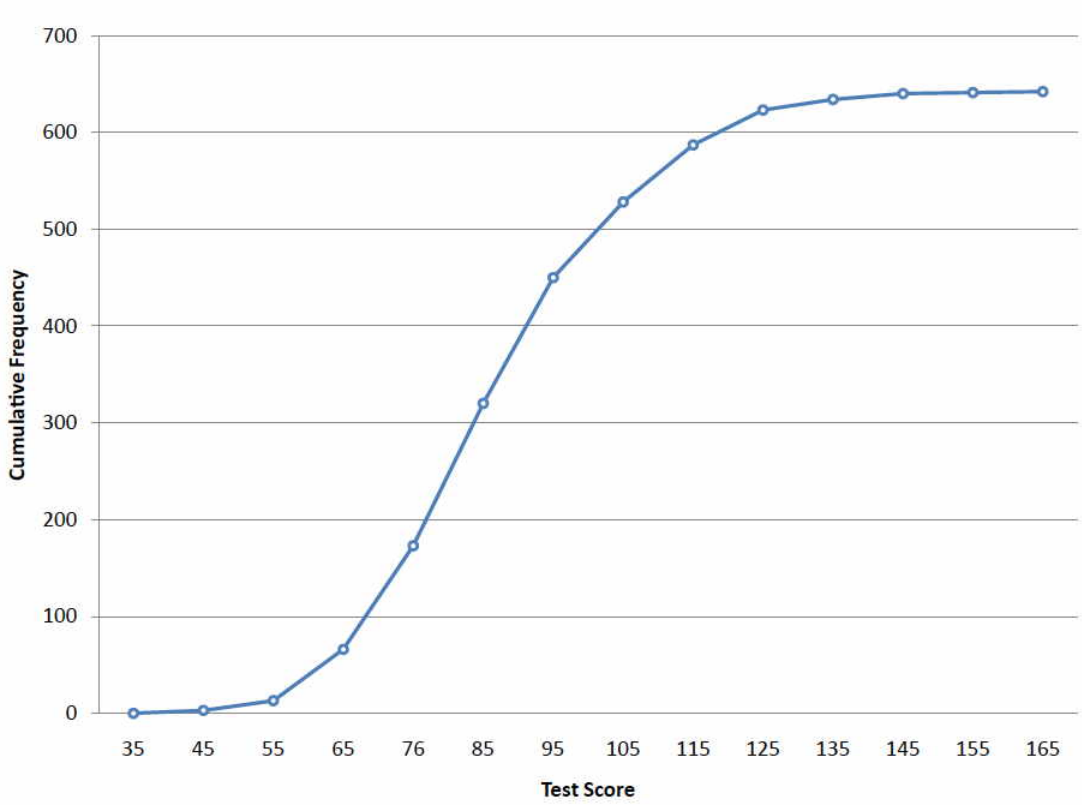


Figure $2.5.2$: Cumulative frequency polygon for the psychology test scores

A cumulative frequency polygon for the same test scores is shown in Figure $2.5.2$. The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals. For example, there are no scores in the interval labeled 35, three in the interval 45, and 10 in the interval 55. Therefore, the Y value corresponding to "55" is 13. Since 642 students took the test, the cumulative frequency for the last interval is 642.

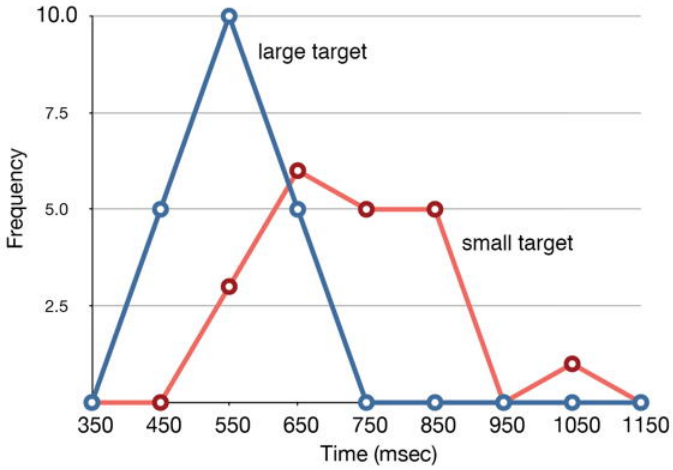


Figure 2.5.3: Overlaid frequency polygons

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 2.5.3 provides an example. The data come from a task in which the goal is to move a computer cursor to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together in Figure 2.5.3. The figure shows that, although there is some overlap in times, it generally took longer to move the cursor to the small target than to the large one.

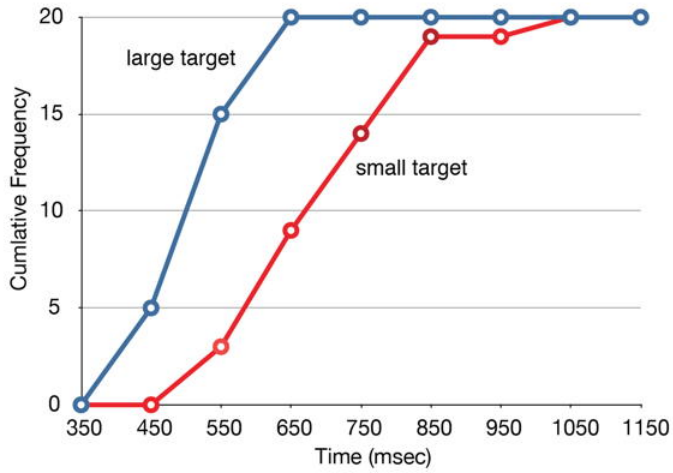


Figure 2.5.1: Overlaid cumulative frequency polygons

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in Figure 2.5.4 using the same data from the cursor task. The difference in distributions for the two targets is again evident.

2.6: Box Plots

Learning Objectives

- Define basic terms including hinges, H-spread, step, adjacent value, outside value, and far out value
- Create a box plot
- Create parallel box plots
- Determine whether a box plot is appropriate for a given data set

We have already discussed techniques for visually representing data (see histograms and frequency polygons). In this section, we present another important graph called a box plot. Box plots are useful for identifying outliers and for comparing distributions. We will explain box plots with the help of data from an in-class experiment. As part of the "Stroop Interference Case Study," students in introductory statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. We'll compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender. Such a display is said to involve parallel box plots.

There are several steps in constructing a box plot. The first relies on the 25th, 50th, $and\ 75^{th}$ percentiles in the distribution of scores. Figure $2.6.1$shows how these three statistics are used. For each gender, we draw a box extending from the 25th percentile to the 75th percentile. The 50th percentile is drawn inside the box. Therefore,

- the bottom of each box is the 25th percentile,
- the top is the 75th percentile,
- and the line in the middle is the 50th percentile.

The data for the women in our sample are shown in Table $2.6.1$.

Table $2.6.1$: Women's times

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

For these data, the 25th percentile is 17, the 50th percentile is 19, and the 75th percentile is 20. For the men (whose data are not shown), the 25th percentile is 19, the 50th percentile is 22.5, and the 75th percentile is 25.5.

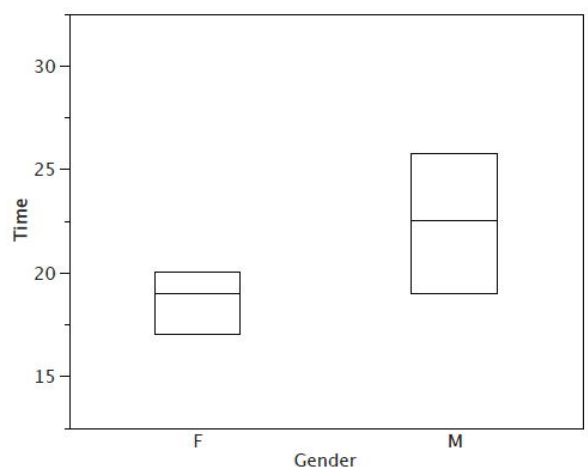


Figure $2.6.1$: The first step in creating box plots

Before proceeding, the terminology in Table $2.6.2$is helpful.

Table $2.6.2$: Box plot terms and values for women's times

Name	Formula	Value
Upper Hinge	75th Percentile	20
Lower Hinge	25th Percentile	17
H-Spread	Upper Hinge - Lower Hinge	3
Step	1.5 x H-Spread	4.5
Upper Inner Fence	Upper Hinge + 1 Step	24.5
Lower Inner Fence	Lower Hinge - 1 Step	12.5
Upper Outer Fence	Upper Hinge + 2 Steps	29
Lower Outer Fence	Lower Hinge - 2 Steps	8
Upper Adjacent	Largest value below Upper Inner Fence	24
Lower Adjacent	Smallest value above Lower Inner Fence	14
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence	29
Far Out Value	A value beyond an Outer Fence	None

Continuing with the box plots, we put "whiskers" above and below each box to give additional information about the spread of the data. Whiskers are vertical lines that end in a horizontal stroke. Whiskers are drawn from the upper and lower hinges to the upper and lower adjacent values (24 and 14 for the women's data).

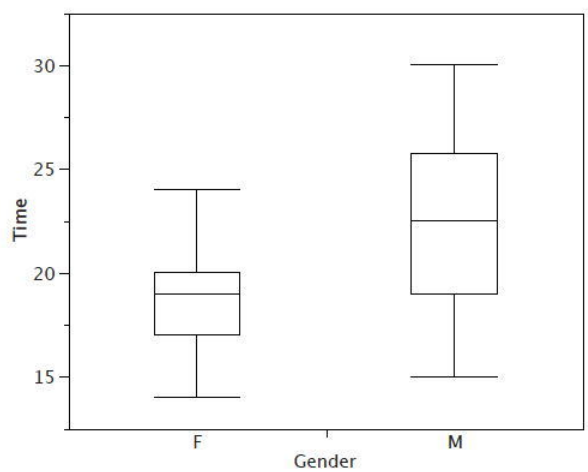


Figure $2.6.2$: The box plots with the whiskers drawn

Although we don't draw whiskers all the way to outside or far out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small "$o's$" and far out values are indicated by asterisks ($*$). In our data, there are no far out values and just one outside value. This outside value of 29 is for the women and is shown in Figure $2.6.3$.

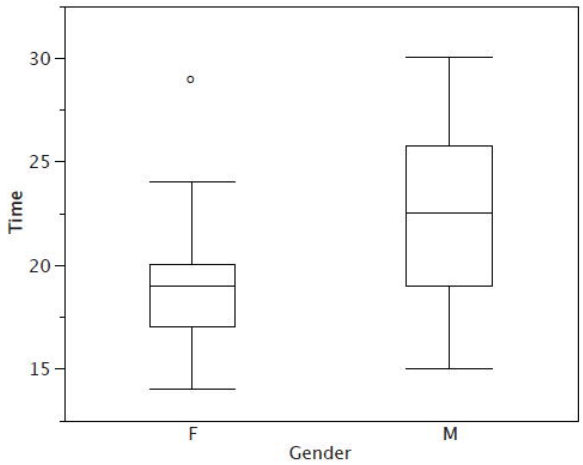


Figure 2.6.3: The box plots with the outside value shown

There is one more mark to include in box plots (although sometimes it is omitted). We indicate the mean score for a group by inserting a plus sign. Figure 2.6.4 shows the result of adding means to our box plots.

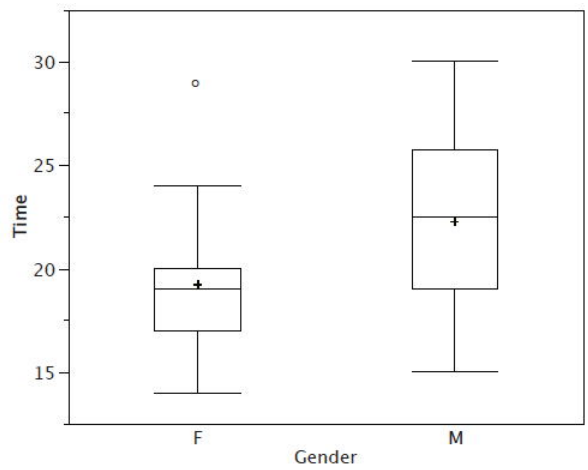


Figure 2.6.4: The completed box plots

Figure 2.6.4 provides a revealing summary of the data. Since half the scores in a distribution are between the hinges (recall that the hinges are the 25th and 75th percentiles), we see that half the women's times are between 17 and 20 seconds, whereas half the men's times are between 19 and 25.5. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men. Figure 2.6.5 shows the box plot for the women's data with detailed labels.

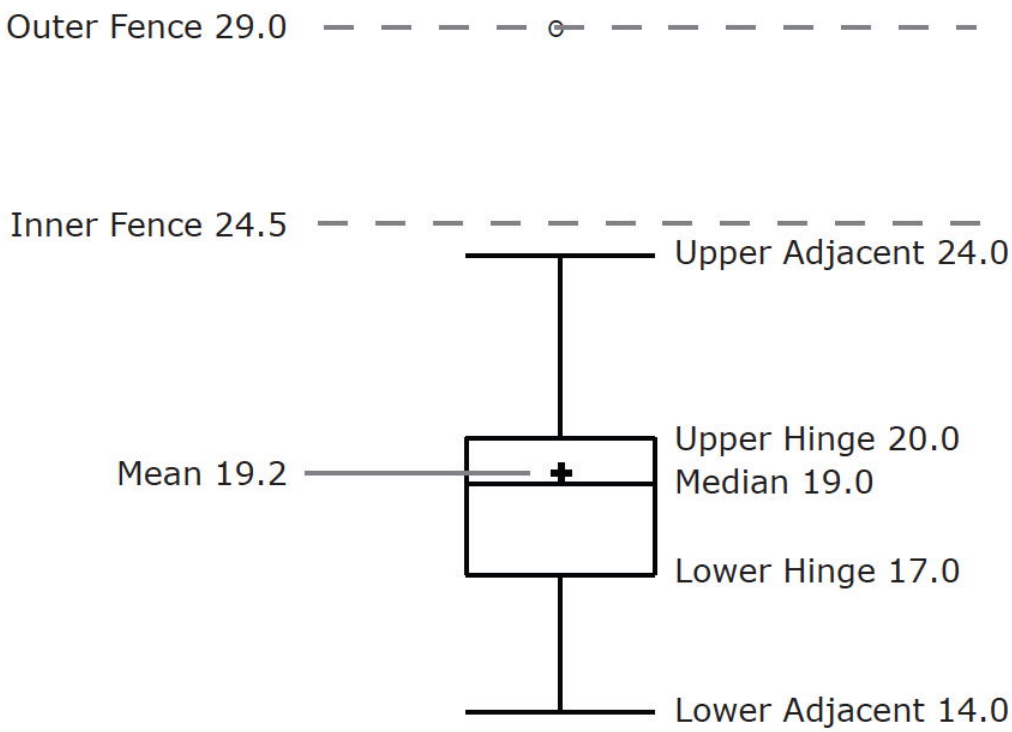


Figure 2.6.5: The box plot for the women's data with detailed labels

Box plots provide basic information about a distribution. For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. A larger mean than median would also indicate a positive skew. Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot, and to examine these details one should create a histogram and/or a stem and leaf display.

Here are some other examples of box plots:

Example 2.6.1: Time to move the mouse over a target

The data come from a task in which the goal is to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The box plots of the two distributions are shown below. You can see that although there is some overlap in times, it generally took longer to move the mouse to the small target than to the large one.

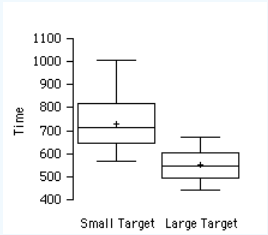


Figure 2.6.6: Target Box Plots

Example 2.6.2: Draft lottery

In 1969 the war in Vietnam was at its height. An agency called the *Selective Service* was charged with finding a fair procedure to determine which young men would be conscripted ("drafted") into the U.S. military. The procedure was supposed to be fair in the sense of not favoring any culturally or economically defined subgroup of American men. It was decided that choosing "draftees" solely on the basis of a person's birth date would be fair. A birthday lottery was thus devised. Pieces of paper representing the 366 days of the year (including February 29) were placed in plastic capsules, poured into a rotating drum,

and then selected one at a time. The lower the draft number, the sooner the person would be drafted. Men with high enough numbers were not drafted at all.

The first number selected was 258, which meant that someone born on the 258th day of the year (September 14) would be among the first to be drafted. The second number was 115, so someone born on the 115th day (April 24) was among the second group to be drafted. All 366 birth dates were assigned draft numbers in this way.

To crate box plots, we divided the 366 days of the year into thirds. The first third goes from $\text{January } 1 \text{ to May } 1$, the second from $\text{May } 2 \text{ to August } 31$, and the last from $\text{September } 1 \text{ to December } 31$. The three groups of birth dates yield three groups of draft numbers. The draft number for each birthday is the order it was picked in the drawing. The figure below contains box plots of the three sets of draft numbers. As you can see, people born later in the year tended to have lower draft numbers.

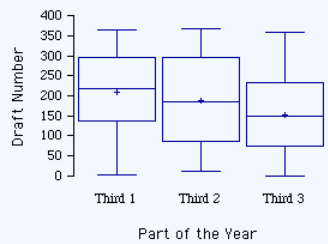


Figure $2.6.7$: Draft Lottery Box Plots

Variations on box plots

Statistical analysis programs may offer options on how box plots are created. For example, the box plots in Figure $2.6.8$ are constructed from our data but differ from the previous box plots in several ways.

1. It does not mark outliers.
2. The means are indicated by green lines rather than plus signs.
3. The mean of all scores is indicated by a gray line.
4. Individual scores are represented by dots. Since the scores have been rounded to the nearest second, any given dot might represent more than one score.
5. The box for the women is wider than the box for the men because the widths of the boxes are proportional to the number of subjects of each gender (31 women and 16 men).

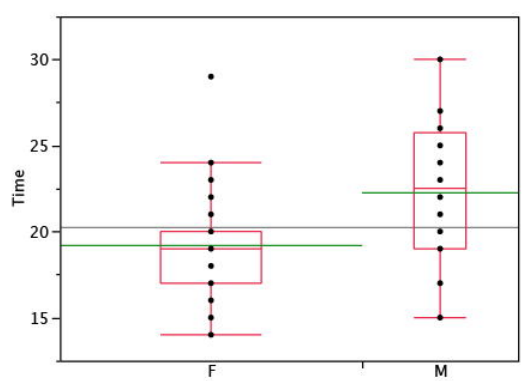


Figure $2.6.8$: Box plots showing the individual scores and the means

Each dot in Figure $2.6.8$ represents a group of subjects with the same score (rounded to the nearest second). An alternative graphing technique is to jitter thepoints. This means spreading out different dots at the same horizontal position, one dot for each subject. The exact horizontal position of a dot is determined randomly (under the constraint that different dots don't overlap exactly). Spreading out the dots helps you to see multiple occurrences of a given score. However, depending on the dot size and the screen resolution, some points may be obscured even if the points are jittererd. Figure $2.6.9$ shows what jittering looks like.

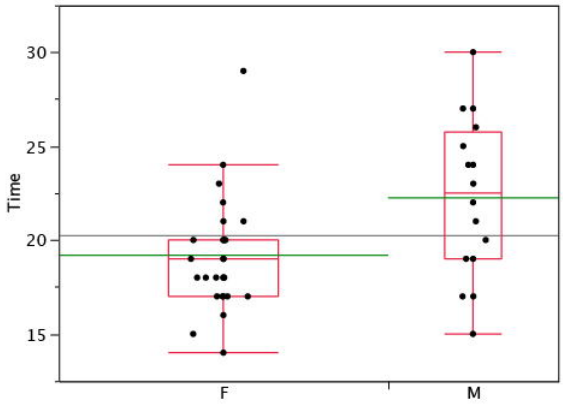


Figure 2.6.9: Box plots with the individual scores jittered

Different styles of box plots are best for different situations, and there are no firm rules for which to use. When exploring your data, you should try several ways of visualizing them. Which graphs you include in your report should depend on how well different graphs reveal the aspects of the data you consider most important.

2.7: Box Plot Demo

Learning Objectives

- Understand what it means for a distribution to balance on a fulcrum
- Learn which measure of central tendency will balance a distribution

Instructions

- The box plot in the simulation is based on the data shown to the left of the box plot.
- Notice that various aspects of the box plot such as the mean and median are labeled.
  - These labels can be hidden by unchecking the "show labels on box plot" button.
- Beneath the data is a pair of buttons that let you specify whether you want to be able to enter data or to see statistics based on the data.
- When you use the simulation, try to modify the data in various ways and see how it affects the box plot.
  - Try putting in some extreme values and see if they get labeled as outliers.
  - Outside values are shown as "$o's$" beyond the inner fence.
  - Far out values are shown as $*'s$ outside the outer fence.
- You can delete all the data (by pressing the "Clear All button") and enter your own data.
  - Your data can be typed in or pasted in from another application.
  - When pasting, use keyboard shortcut for pasting (Command-V for Mac, CTRL-V for Windows).
  - When you change the data, the box plot will disappear.

After you have entered new data, click the "Draw box plot" button to redraw the box plot.

Illustrated Instructions

The screenshot below shows the box plot simulation with its default data on the left.

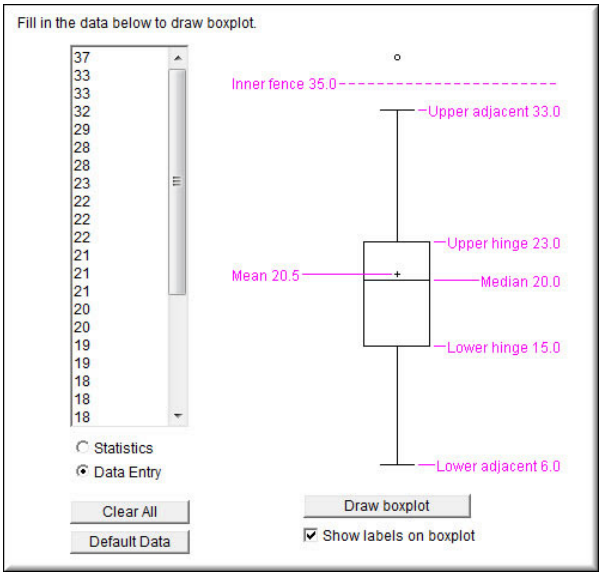


Figure 2.7.1: Box Plot simulation

You can change some or all of the data on the left by editing individual numbers or pasting in your own data. When pasting, you must use the keyboard shortcut for pasting (Command-V for Mac, CTRL-V for Windows).

2.8: Bar Charts

Learning Objectives

- Create and interpret bar charts
- Judge whether a bar chart or another graph such as a box plot would be more appropriate

In the section on qualitative variables, we saw how bar charts could be used to illustrate the frequencies of different categories. For example, the bar chart shown in Figure 2.8.1 shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

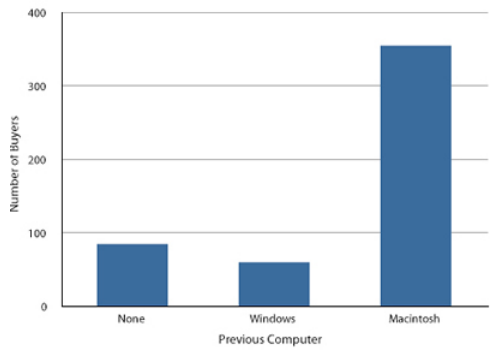


Figure 2.8.1: iMac buyers as a function of previous computer ownership

In this section, we show how bar charts can be used to present other kinds of quantitative information, not just frequency counts. The bar chart in Figure 2.8.2 shows the percent increases in the Dow Jones, Standard and Poor 500 (S & P), and Nasdaq stock indexes from May 24 2000 to May 24 2001. Notice that both the S & P and the Nasdaq had "negative increases" which means that they decreased in value. In this bar chart, the Y-axis is not frequency but rather the signed quantity *percentage increase.*

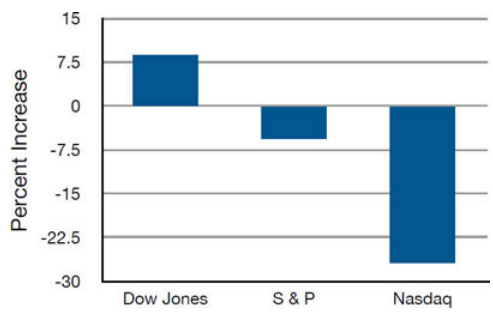


Figure 2.8.2: Percent increase in three stock indexes from May 24th 2000 to May 24th 2001

Bar charts are particularly effective for showing change over time. Figure 2.8.3, for example, shows the percent increase in the Consumer Price Index (CPI) over four three-month periods. The fluctuation in inflation is apparent in the graph.

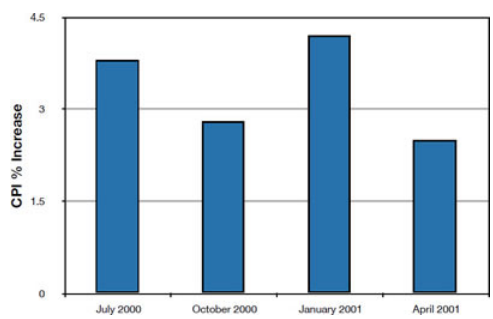


Figure 2.8.3: Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated

Bar charts are often used to compare the means of different experimental conditions. Figure 2.8.4 shows the mean time it took one of us (DL) to move the mouse to either a small target or a large target. On average, more time was required for small targets than for large ones.

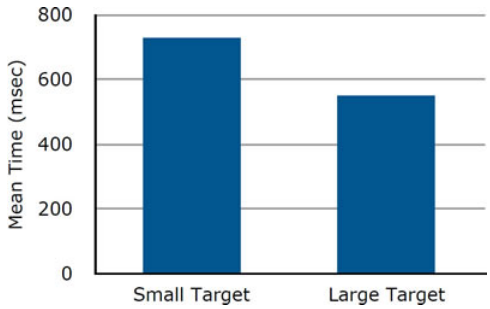


Figure 2.8.4: Bar chart showing the means for the two conditions

Although bar charts can display means, we do not recommend them for this purpose. Box plots should be used instead since they provide more information than bar charts without taking up more space. For example, a box plot of the mouse-movement data is shown in Figure 2.8.5. You can see that Figure 2.8.5 reveals more about the distribution of movement times than does Figure 2.8.4.

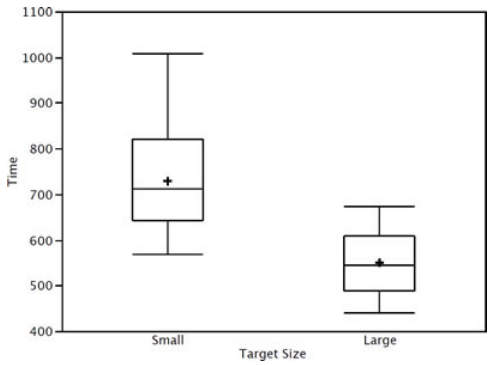


Figure 2.8.5: Box plots of times to move the mouse to the small and large targets

The section on qualitative variables presented earlier in this chapter discussed the use of bar charts for comparing distributions. Some common graphical mistakes were also noted. The earlier discussion applies equally well to the use of bar charts to display quantitative variables.

2.9: Line Graphs

Learning Objectives

- Create and interpret line graphs
- Judge whether a line graph would be appropriate for a given data set

A line graph is a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). For example, Figure $2.9.1$ was presented in the section on bar charts and shows changes in the Consumer Price Index (CPI) over time.

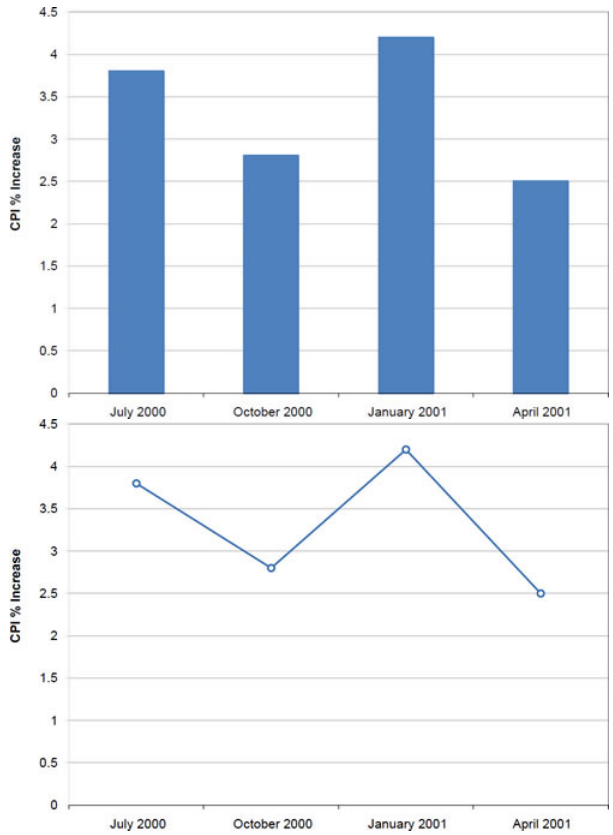


Figure $2.9.2$: A line graph of the percent change in the CPI over time. Each point represents percent increase for the three months ending at the date indicated.

A line graph of these same data is shown in Figure $2.9.2$. Although the figures are similar, the line graph emphasizes the change from period to period.

Line graphs are appropriate only when both the X- and Y-axes display ordered (rather than qualitative) variables. Although bar graphs can also be used in this situation, line graphs are generally better at comparing changes over time. Figure $2.9.3$, for example, shows percent increases and decreases in five components of the Consumer Price Index (CPI). The figure makes it easy to see that medical costs had a steadier progression than the other components. Although you could create an analogous bar chart, its interpretation would not be as easy.

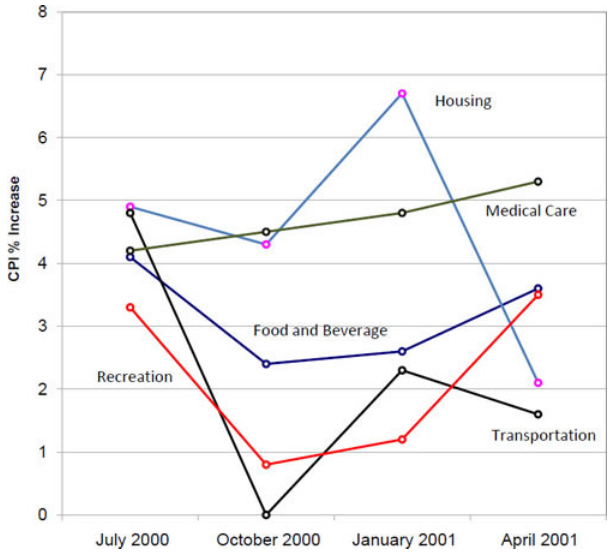


Figure 2.9.3: A line graph of the percent change in five components of the CPI over time.

Let us stress that it is misleading to use a line graph when the X-axis contains merely qualitative variables. Figure 2.9.4 inappropriately shows a line graph of the card game data from Yahoo, discussed in the section on qualitative variables. The defect in Figure 2.9.4 is that it gives the false impression that the games are naturally ordered in a numerical way.

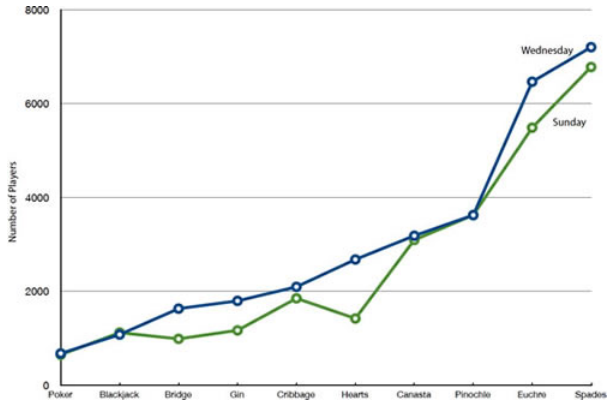


Figure 2.9.4: A line graph, inappropriately used, depicting the number of people playing different card games on Sunday and Wednesday.

2.10: Dot Plots

Learning Objectives

- Create and interpret dot plots
- Judge whether a dot plot would be appropriate for a given data set

Dot plots can be used to display various types of information. Figure 2.10.1 uses a dot plot to display the number of M & M's of each color found in a bag of M & M's. Each dot represents a single M & M. From the figure, you can see that there were 3 blue M & M's, 19 brown M & M's, etc.

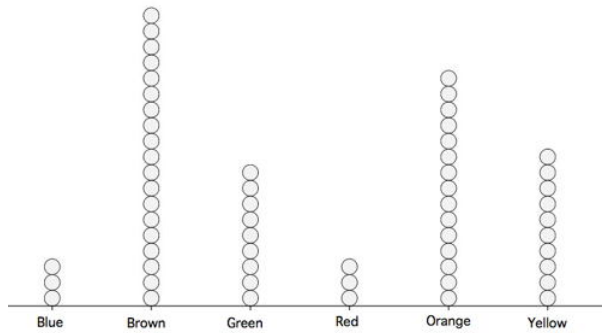


Figure 2.10.1: A dot plot showing the number of M & M's of various colors in a bag of M & M's

The dot plot in Figure 2.10.2 shows the number of people playing various card games on the Yahoo website on a Wednesday. Unlike Figure 2.10.1, the location rather than the number of dots represents the frequency.

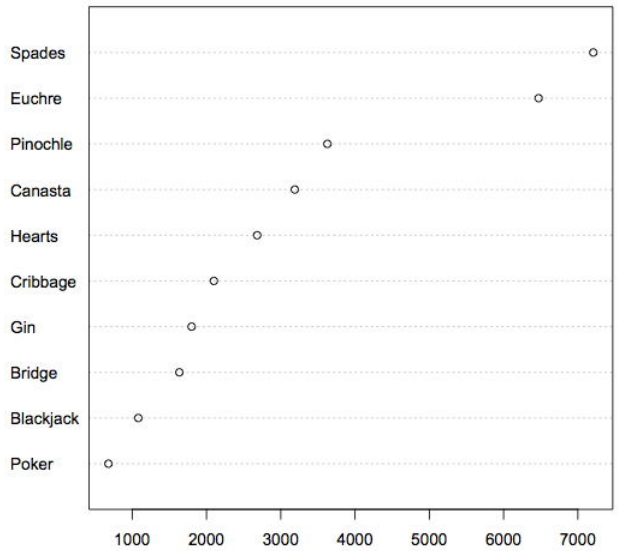


Figure 2.10.2: A dot plot showing the number of people playing various card games on a Wednesday

The dot plot in Figure 2.10.3 shows the number of people playing on a Sunday and on a Wednesday. This graph makes it easy to compare the popularity of the games separately for the two days, but does not make it easy to compare the popularity of a given game on the two days.

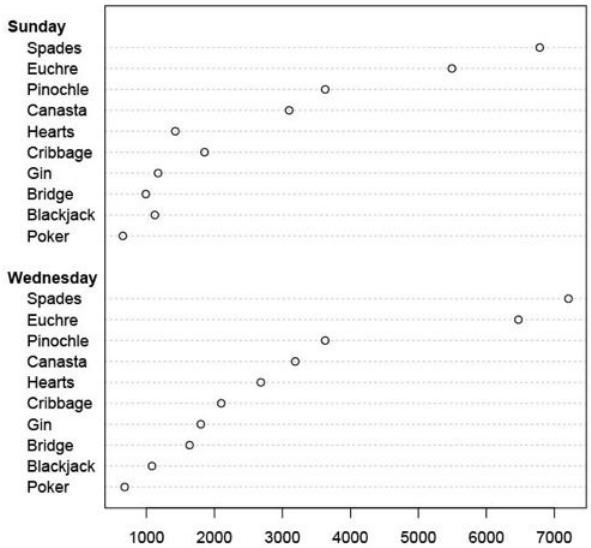


Figure 2.10.3: A dot plot showing the number of people playing various card games on a Sunday and on a Wednesday

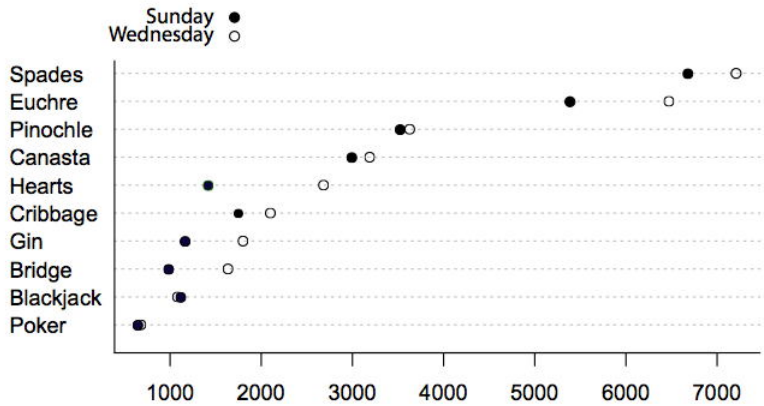


Figure 2.10.4: An alternate way of showing the number of people playing various card games on a Sunday and on a Wednesday

The dot plot in Figure 2.10.4 makes it easy to compare the days of the week for specific games while still portraying differences among games.

2.11: Statistical Literacy

Are Commercial Vehicles in Texas Unsafe?

Prerequisites

Graphing Distributions

A news report on the safety of commercial vehicles in Texas stated that one out of five commercial vehicles have been pulled off the road in 2012 because they were unsafe. In addition, $12,301$ commercial drivers have been banned from the road for safety violations.

The author presents the bar chart below to provide information about the percentage of fatal crashes involving commercial vehicles in Texas since 2006. The author also quotes DPS director Steven McCraw:

> Commercial vehicles are responsible for approximately 15 percent of the fatalities in Texas crashes. Those who choose to drive unsafe commercial vehicles or drive a commercial vehicle unsafely pose a serious threat to the motoring public.

Example 2.11.1

Based on what you have learned in this chapter, does this bar chart below provide enough information to conclude that unsafe or unsafely driven commercial vehicles pose a serious threat to the motoring public? What might you conclude if 30 percent of all the vehicles on the roads of Texas in 2010 were commercial and accounted for 16 percent of fatal crashes?

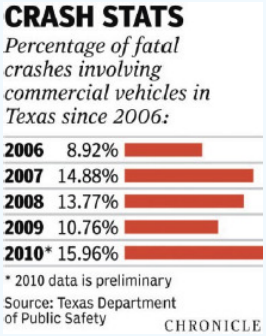


Figure 2.11.1: Crash Statistics for commercial vehicles in Texas

Solution

This bar chart does not provide enough information to draw such a conclusion because we don't know, on the average, in a given year what percentage of all vehicles on the road are commercial vehicles. For example, if 30 percent of all the vehicles on the roads of Texas in 2010 are commercial ones and only 16 percent of fatal crashes involved commercial vehicles, then commercial vehicles are safer than non-commercial ones. Note that in this case 70 percent of vehicles are non-commercial and they are responsible for 84 percent of the fatal crashes.

Linear By Design

Example 2.11.2

Fox News aired the line graph below showing the number unemployed during four quarters between 2007 and 2010.

Figure 2.11.2: Fox news graph showing job loss by quarter

Does Fox News' line graph provide misleading information? Why or Why not?

Solution:

There are major flaws with the Fox News graph. First, the title of the graph is misleading. Although the data show the number unemployed, Fox News' graph is titled "Job **Loss** by Quarter." Second, the intervals on the X-axis are misleading. Although there are 6 months between September 2008 and March 2009 and 15 months between March 2009 and June 2010, the intervals are represented in the graph by very similar lengths. This gives the false impression that unemployment increased steadily.

The graph presented below is corrected so that distances on the X-axis are proportional to the number of days between the dates. This graph shows clearly that the rate of increase in the number unemployed is greater between September 2008 and March 2009 than it is between March 2009 and June 2010.

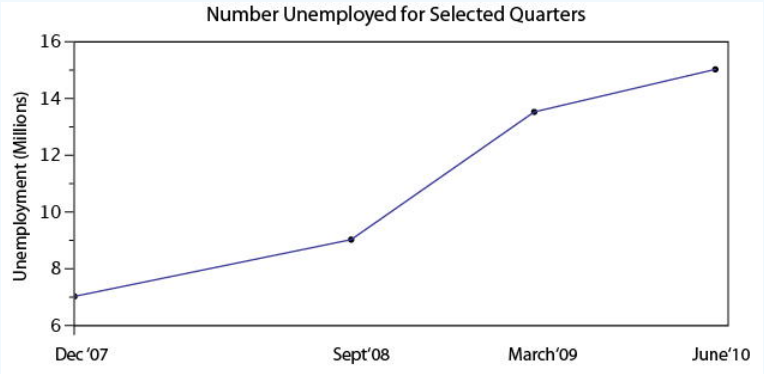


Figure 2.11.3: Corrected Fox News graph

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- Seyd Ercan and David Lane

2.E: Graphing Distributions (Exercises)

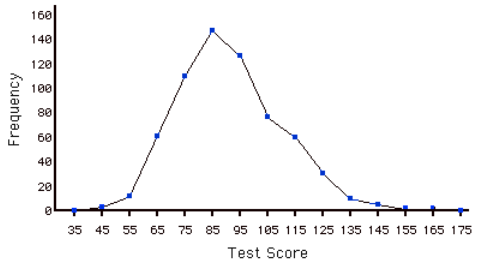
General Questions

Q1

Name some ways to graph quantitative variables and some ways to graph qualitative variables. (relevant section & relevant section)

Q2

Based on the frequency polygon displayed below, the most common test grade was around what score? Explain. (relevant section)



Q3

An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the total number of pieces correctly remembered from **three** chess positions. Create side-by-side box plots for these three groups. What can you say about the differences between these groups from the box plots? (relevant section)

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

Q4

You have to decide between displaying your data with a histogram or with a stem and leaf display. What factor(s) would affect your choice? (relevant section & relevant section)

Q5

In a box plot, what percent of the scores are between the lower and upper hinges? (relevant section)

Q6

A student has decided to display the results of his project on the number of hours people in various countries slept per night. He compared the sleeping patterns of people from the US, Brazil, France, Turkey, China, Egypt, Canada, Norway, and Spain. He was

planning on using a line graph to display this data. Is a line graph appropriate? What might be a better choice for a graph? (relevant section & relevant section)

Q7

For the data from the 1977 Stat. and Biom. 200 class for eye color, construct: (relevant section)

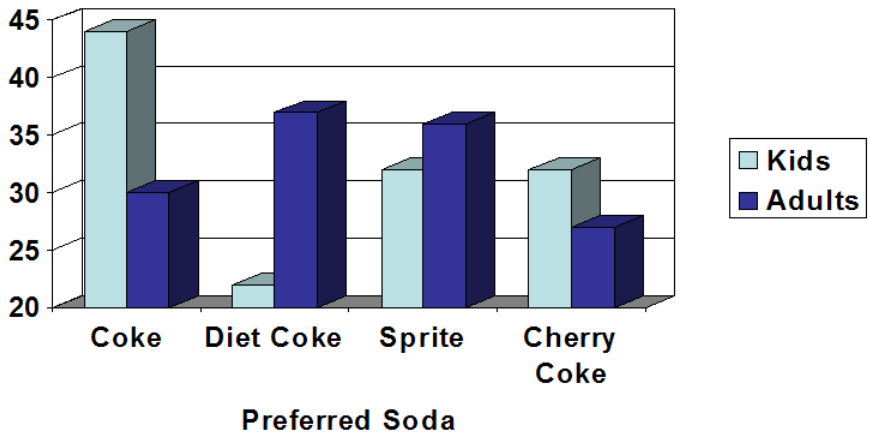
a. pie graph
b. horizontal bar graph
c. vertical bar graph
d. a frequency table with the relative frequency of each eye color

Eye Color	Number of students
Brown	11
Blue	10
Green	4
Gray	1

(Question submitted by J. Warren, UNH)

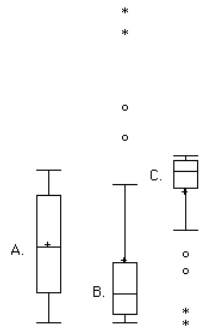
Q8

A graph appears below showing the number of adults and children who prefer each type of soda. There were 130 adults and kids surveyed. Discuss some ways in which the graph below could be improved. (relevant section)



Q9

Which of the box plots below has a large positive skew? Which has a large negative skew? (relevant section & relevant section)



Questions from Case Studies

The following questions are from the Angry Moods (AM) case study.

Q10

(AM#6) Is there a difference in how much males and females use aggressive behavior to improve an angry mood? For the "Anger-Out" scores:

a. Create parallel box plots. (relevant section)
b. Create a back to back stem and leaf displays (You may have trouble finding a computer to do this so you may have to do it by hand.) (relevant section)

Q11

(AM#9) Create parallel box plots for the Anger-In scores by sports participation. (relevant section)

Q12

(AM#11) Plot a histogram of the distribution of the Control-Out scores. (relevant section)

Q13

(AM#14) Create a bar graph comparing the mean Control-In score for the athletes and the non-athletes. What would be a better way to display this data? (relevant section)

Q14

(AM#18) Plot parallel box plots of the Anger Expression Index by sports participation. Does it look like there are any outliers? Which group reported expressing more anger? (relevant section)

The following questions are from the Flatulence (F) case study.

Q15

(F#1) Plot a histogram of the variable "per day." (relevant section)

Q16

(F#7) Create parallel box plots of "how long" as a function gender. Why is the 25th percentile not showing? What can you say about the results? (relevant section)

Q17

(F#9) Create a stem and leaf plot of the variable "how long" What can you say about the shape of the distribution? (relevant section.1)

The following questions are from the Physicians' Reactions (PR) case study.

Q18

(PR#1) Create box plots comparing the time expected to be spent with the average-weight and overweight patients. (relevant section)

Q19

(PR#4) Plot histograms of the time spent with the average-weight and overweight patients. (relevant section)

Q20

(PR#5) To which group does the patient with the highest expected time belong?

The following questions are from the Smiles and Leniency (SL) case study

Q21

(SL#1) Create parallel box plots for the four conditions. (relevant section)

Q22

(SL#3) Create back to back stem and leaf displays for the false smile and neutral conditions. (It may be hard to find a computer program to do this for you, so be prepared to do it by hand). (relevant section)

The following questions are from the ADHD Treatment (AT) case study.

Q23

(AT#3) Create a line graph of the data. Do certain dosages appear to be more effective than others? (relevant section)

Q24

(AT#5) Create a stem and leaf plot of the number of correct responses of the participants after taking the placebo ($d0$ variable). What can you say about the shape of the distribution? (relevant section)

Q25

Create box plots for the four conditions. You may have to rearrange the data to get a computer program to create the box plots.

The following question is from the SAT and College GPA case study.

Q26

Create histograms and stem and leaf displays of both high-school grade point average and university grade point average. In what way(s) do the distributions differ?

Q27

The April 10th issue of the Journal of the American Medical Association reports a study on the effects of anti-depressants. The study involved 340 subjects who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort (an herb), Zoloft (Pfizer's cousin of Lilly's Prozac) or placebo for an 8-week period. The following are the mean scores (approximately) for the three groups of subjects over the eight-week experiment. The first column is the baseline. Lower scores mean less depression. Create a graph to display these means.

Placebo	22.5	19.1	17.9	17.1	16.2	15.1	12.1	12.3
Wort	23.0	20.2	18.2	18.0	16.5	16.1	14.2	13.0
Zoloft	22.4	19.2	16.6	15.5	14.2	13.1	11.8	10.5

The following questions are from Assessment Resource Tools for Improving Statistical Thinking **. Visit the site**

Q28

For the graph below, of heights of singers in a large chorus, please write a complete description of the histogram. Be sure to comment on all the important features.

Q29

Pretend you are constructing a histogram for describing the distribution of salaries for individuals who are 40 years or older, but are not yet retired.

a. What is on the Y-axis? Explain.
b. What is on the X-axis?
c. What would be the probable shape of the salary distribution? Explain why.

Select Answers

S2

85

CHAPTER OVERVIEW

3: Summarizing Distributions

Descriptive statistics often involves using a few numbers to summarize a distribution. One important aspect of a distribution is where its center is located. Measures of central tendency are discussed first. A second aspect of a distribution is how spread out it is. In other words, how much the numbers in the distribution vary from one another. The second section describes measures of variability. Distributions can differ in shape. Some distributions are symmetric whereas others have long tails in just one direction. The third section describes measures of the shape of distributions. The final two sections concern (1) how transformations affect measures summarizing distributions and (2) the variance sum law, an important relationship involving a measure of variability.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

3.1: Central Tendency

Central tendency is a loosely defined concept that has to do with the location of the center of a distribution. The section "What is Central Tendency" presents three definitions of the center of a distribution. "Measures of Central Tendency" presents the three most common measures of the center of the distribution. The three simulations that follow relate the definitions of the center of a distribution to the commonly used measures of central tendency. The findings from these simulations are summarized in the section "Mean and Median." The "Mean and Median" allows you to explore how the relative size of the mean and the median depends on the skew of the distribution.

Less frequently used measures of central tendency can be valuable supplements to the more commonly used measures. Some of these measures are presented in "Additional Measures." Finally, the last section compares and summarizes differences among measures of central tendency.

3.2: What is Central Tendency

Learning Objectives

- Identify situations in which knowing the center of a distribution would be valuable
- Give three different ways the center of a distribution can be defined
- Describe how the balance is different for symmetric distributions than it is for asymmetric distributions

What is "central tendency," and why do we want to know the central tendency of a group of scores? Let us first try to answer these questions intuitively. Then we will proceed to a more formal discussion.

Imagine this situation: You are in a class with just four other students, and the five of you took a 5-point pop quiz. Today your instructor is walking around the room, handing back the quizzes. She stops at your desk and hands you your paper. Written in bold black ink on the front is "$3/5$." How do you react? Are you happy with your score of 3 or disappointed? How do you decide? You might calculate your percentage correct, realize it is 60%, and be appalled. But it is more likely that when deciding how to react to your performance, you will want additional information. What additional information would you like?

If you are like most students, you will immediately ask your neighbors, "Whad'ja get?" and then ask the instructor, "How did the class do?" In other words, the additional information you want is how your quiz score compares to other students' scores. You therefore understand the importance of comparing your score to the class distribution of scores. Should your score of 3 turn out to be among the higher scores, then you'll be pleased after all. On the other hand, if 3 is among the lower scores in the class, you won't be quite so happy.

This idea of comparing individual scores to a distribution of scores is fundamental to statistics. So let's explore it further, using the same example (the pop quiz you took with your four classmates). Three possible outcomes are shown in Table $3.2.1$. They are labeled "Dataset A," "Dataset B," and "Dataset C." Which of the three datasets would make you happiest? In other words, in comparing your score with your fellow students' scores, in which dataset would your score of 3 be the most impressive?

In Dataset A, everyone's score is 3. This puts your score at the exact center of the distribution. You can draw satisfaction from the fact that you did as well as everyone else. But of course it cuts both ways: everyone else did just as well as you.

Table $3.2.1$: Three possible datasets for the 5-point make-up quiz

Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

Now consider the possibility that the scores are described as in Dataset B. This is a depressing outcome even though your score is no different than the one in Dataset A. The problem is that the other four students had higher grades, putting yours below the center of the distribution.

Finally, let's look at Dataset C. This is more like it! All of your classmates score lower than you so your score is above the center of the distribution.

Now let's change the example in order to develop more insight into the center of a distribution. Figure $3.2.1$ shows the results of an experiment on memory for chess positions. Subjects were shown a chess position and then asked to reconstruct it on an empty chess board. The number of pieces correctly placed was recorded. This was repeated for two more chess positions. The scores represent the total number of chess pieces correctly placed for the three chess positions. The maximum possible score was 89.

Two groups are compared. On the left are people who don't play chess. On the right are people who play a great deal (tournament players). It is clear that the location of the center of the distribution for the non-players is much lower than the center of the distribution for the tournament players.

	8	05
	7	156
	6	233
	5	168
330	4	06
9420	3	
622	2	

Figure 3.2.1: Back-to-back stem and leaf display. The left side shows the memory scores of the non-players. The right side shows the scores of the tournament players.

We're sure you get the idea now about the center of a distribution. It is time to move beyond intuition. We need a formal definition of the center of a distribution. In fact, we'll offer you three definitions! This is not just generosity on our part. There turn out to be (at least) three different ways of thinking about the center of a distribution, all of them useful in various contexts. In the remainder of this section we attempt to communicate the idea behind each concept. In the succeeding sections we will give statistical measures for these concepts of central tendency.

Definitions of Center

Now we explain the three different ways of defining the center of a distribution. All three are called measures of central tendency.

Balance Scale

One definition of central tendency is the point at which the distribution is in balance. Figure $3.2.2$ shows the distribution of the five numbers $2, 3, 4, 9, 16$ placed upon a balance scale. If each number weighs one pound, and is placed at its position along the number line, then it would be possible to balance them by placing a fulcrum at 6.8.

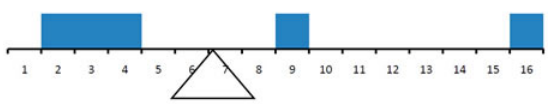


Figure 3.2.2: A balance scale

For another example, consider the distribution shown in Figure $3.2.3$. It is balanced by placing the fulcrum in the geometric middle.

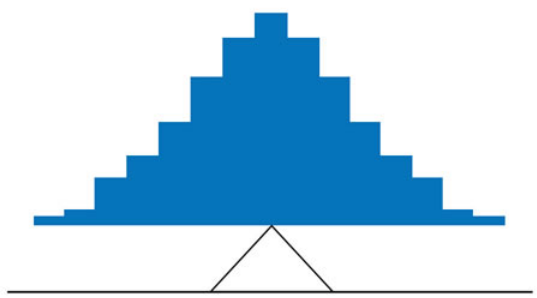


Figure 3.2.3: A distribution balanced on the tip of a triangle.

Figure $3.2.4$ illustrates that the same distribution can't be balanced by placing the fulcrum to the left of center.

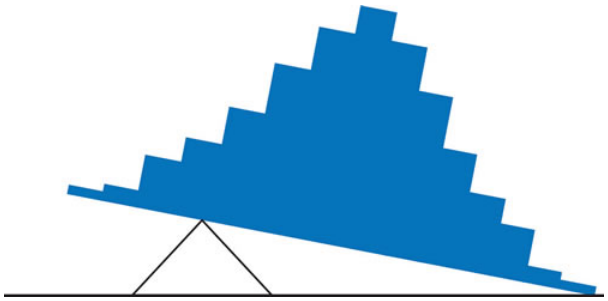


Figure 3.2.4: The distribution is not balanced

Figure 3.2.5 shows an asymmetric distribution. To balance it, we cannot put the fulcrum halfway between the lowest and highest values (as we did in Figure 3.2.3). Placing the fulcrum at the "half way" point would cause it to tip towards the left.

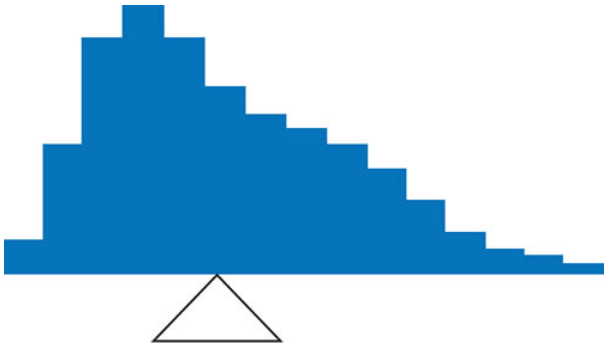


Figure 3.2.5: An asymmetric distribution balanced on the tip of a triangle

The balance point defines one sense of a distribution's center. The simulation in the next section "Balance Scale Simulation" shows how to find the point at which the distribution balances.

Smallest Absolute Deviation

Another way to define the center of a distribution is based on the concept of the sum of the absolute deviations (differences). Consider the distribution made up of the five numbers $2, 3, 4, 9, 16$ Let's see how far the distribution is from 10 (picking a number arbitrarily). Table 3.2.2 shows the sum of the absolute deviations of these numbers from the number 10.

Table 3.2.2: An example of the sum of absolute deviations

Values	Absolute Deviations from 10
2	8
3	7
4	6
9	1
16	6
Sum	28

The first row of the table shows that the absolute value of the difference between 2 and 10 is 8; the second row shows that the absolute difference between 3 and 10 is 7, and similarly for the other rows. When we add up the five absolute deviations, we get 28. So, the sum of the absolute deviations from 10 is 28. Likewise, the sum of the absolute deviations from 5 equals $3+2+1+4+11=21$. So, the sum of the absolute deviations from 5 is smaller than the sum of the absolute deviations from 10. In this sense, 5 is closer, overall, to the other numbers than is 10.

We are now in a position to define a second measure of central tendency, this time in terms of absolute deviations. Specifically, according to our second definition, the center of a distribution is the number for which the sum of the absolute deviations is smallest. As we just saw, the sum of the absolute deviations from 10 is 28 and the sum of the absolute deviations from 5 is 21. Is there a value for which the sum of the absolute deviations is even smaller than 21? Yes. For these data, there is a value for which

the sum of absolute deviations is only 20. See if you can find it. A general method for finding the center of a distribution in the sense of absolute deviations is provided in the simulation "Absolute Differences Simulation."

Smallest Squared Deviation

We shall discuss one more way to define the center of a distribution. It is based on the concept of the sum of squared deviations (differences). Again, consider the distribution of the five numbers $2, 3, 4, 9, 16$ Table $3.2.3$ shows the sum of the squared deviations of these numbers from the number 10.

Table $3.2.3$: An example of the sum of squared deviations

Values	Squared Deviations from 10
2	64
3	49
4	36
9	1
16	36
Sum	186

The first row in the table shows that the squared value of the difference between 2 and 10 is 64; the second row shows that the squared difference between 3 and 10 is 49, and so forth. When we add up all these squared deviations, we get 186. Changing the target from 10 to 5, we calculate the sum of the squared deviations from 5 as $9+4+1+16+121=151$. So, the sum of the squared deviations from 5 is smaller than the sum of the squared deviations from 10. Is there a value for which the sum of the squared deviations is even smaller than 151? Yes, it is possible to reach 134.8. Can you find the target number for which the sum of squared deviations is 134.8?

The target that minimizes the sum of squared deviations provides another useful definition of central tendency (the last one to be discussed in this section). It can be challenging to find the value that minimizes this sum. You will see how you do it in the upcoming section "Squared Differences Simulation."

Contributor

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- David M. Lane and Heidi Ziemer

3.3: Measures of Central Tendency

Learning Objectives

- Compute mode

In the previous section we saw that there are several ways to define central tendency. This section defines the three most common measures of central tendency: the mean, the median, and the mode. The relationships among these measures of central tendency and the definitions given in the previous section will probably not be obvious to you. Rather than just tell you these relationships, we will allow you to discover them in the simulations in the sections that follow. This section gives only the basic definitions of the mean, median and mode. A further discussion of the relative merits and proper applications of these statistics is presented in a later section.

Arithmetic Mean

The arithmetic mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. The symbol "μ" is used for the mean of a population. The symbol "M" is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \frac{\sum X}{N} \tag{3.3.1}$$

where $\sum X$ is the sum of all the numbers in the population and N is the number of numbers in the population.

The formula for M is essentially identical:

$$M = \frac{\sum X}{N} \tag{3.3.2}$$

where $\sum X$ is the sum of all the numbers in the sample and N is the number of numbers in the sample.

As an example, the mean of the numbers $1, 2, 3, 6, 8$is $20/5 = 4$ regardless of whether the numbers constitute the entire population or just a sample from the population.

Table $3.3.1$ shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season.

Table $3.3.1$: *Number of touchdown passes*

$$\begin{array}{cccccccccccc} 37 & 33 & 33 & 32 & 29 & 28 & 28 & 23 & 22 & 22 & 22 & 21 \\ 21 & 21 & 20 & 20 & 19 & 19 & 18 & 18 & 18 & 18 & 16 & 15 \\ 14 & 14 & 14 & 12 & 12 & 9 & 6 & & & & & \end{array} \tag{3.3.3}$$

The mean number of touchdown passes thrown is 20.4516 as shown below.

$$\begin{aligned} \mu &= \sum X/N \\ &= 634/31 \\ &= 20.4516 \end{aligned}$$

Although the arithmetic mean is not the only "mean" (there is also a geometric mean), it is by far the most commonly used. Therefore, if the term "mean" is used without specifying whether it is the arithmetic mean, the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean.

Median

The median is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same number of scores is above the median as below it. For the data in Table $3.3.1$, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. The median can also be thought of as the 50th percentile.

Computation of the Median

When there is an odd number of numbers, the median is simply the middle number. For example, the median of $2, 4$, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers $2, 4, 7, 12$ is $(4+7)/2 = 5.5$. When there are numbers with the same values, then the formula for the third definition of the 50th percentile should be used.

Mode

The mode is the most frequently occurring value. For the data in Table $3.3.2$, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see discussion of continuous variables). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table $3.3.2$ shows a grouped frequency distribution for the target response time data. Since the interval with the highest frequency is $600-700$, the mode is the middle of that interval (650).

Table $3.3.2$: Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

3.4: Balance Scale Simulation

Learning Objectives

- Learn which measure of central tendency will balance a distribution

Instructions

This demonstration allows you to change the shape of a distribution and see the point at which the distribution would balance.

The graph in the right panel is a histogram of 600 scores. The mean and median are equal to 8 and are indicated by small vertical bars on the X-axis The top portion of the bar is in blue and represents the mean. The bottom portion is in pink and represents the median. The mean and median are also shown to the left of the Y-axis.

You can see that the histogram is balanced on the tip of the triangle (the fulcrum). You can change the shape of the histogram by painting with the mouse. Notice that the triangle beneath the X-axis automatically moves to the point where the histogram is balanced. Experiment with different shapes and see if you can determine whether there is a relationship between the mean, median, and/or the mode and the location of the balance point.

Illustrated Instructions

Below is a screen shot of the simulaton's beginning screen. Note that the distribution is balanced on the fulcrum. The mean and median are shown to the left and also as small vertical bars below the X-axis. The mean is in blue and the median is in pink. The next figure illustrates this more clearly.

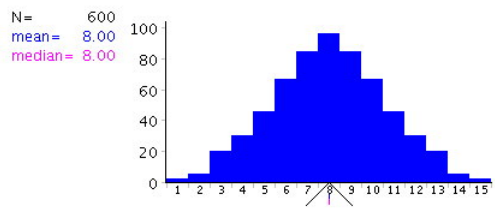


Figure $3.4.1$: Beginning Screen of the Simulation

You can change the distribution by painting with the mouse when running the simulation.Below is an example of the distribution after it has been changed. Note that the mean and median are marked by vertical lines.

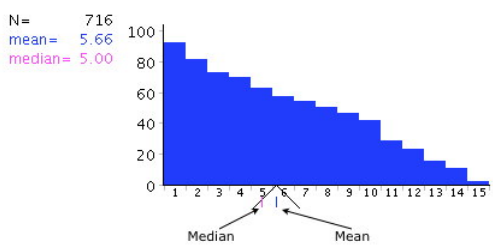


Figure $3.4.2$: Screen of the Simulation after change

3.5: Absolute Differences Simulation

Learning Objectives

- Learn which measure of central tendency minimizes the sum of absolute deviations

Instructions

This demonstration allows you to examine the sum of absolute deviations from a given value. The graph to the right shows the numbers $1, 2, 3, 4,$ and 5 and their deviations from an arbitrary starting value of 0.254 (the figure displays this rounded to 0.25).

The first number, 1, is represented by a red dot. The deviation from 0.254 is represented by a red line from the red dot to the black line. The value of the black line is 0.254. Similarly, the number 2 is represented by a blue dot and its deviation from 0.254 is represented by the length of a blue line.

The portion of the graph with the colored rectangles shows the sum of the absolute deviations. The sum of the deviations is equal to $0.746 + 1.746 + 2.746 + 3.746 + 4.746 = 13.73$ as shown by the height of the colored bar.

In this demonstration, you can move the black bar by clicking on it and dragging it up or down. To see how it works, move it up to 1.0. The deviation of the red point from the black bar is now 0 since they are both 1. The sum of the deviations is now 10.

As you move the bar up and down, the value of the sum of absolute deviations changes. See if you can find the placement of the black bar that produces the smallest value for the sum of the absolute deviations. To check and see if you found the smallest value, click the "OK" button at the bottom of the graph. It will move the bar to the location that produces the smallest sum of absolute deviations.

You can also move the individual points. Click on one of the points and move it up or down and note the effect. Your goal for this demonstration is to discover a rule for determining what value will give you the smallest sum of absolute deviations. When you have discovered the rule, go back and answer the questions again.

Illustrated Instructions

Below is a screen shot of the simulaton's beginning screen.

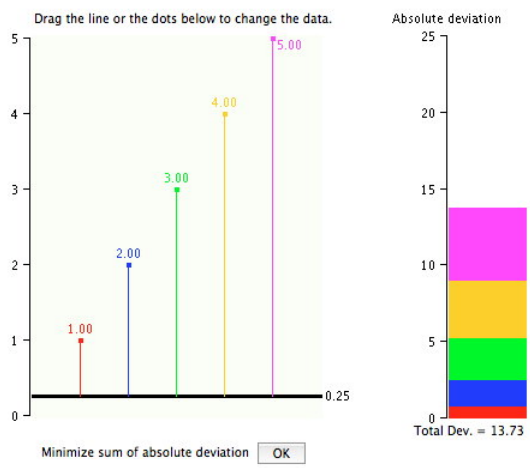


Figure $3.5.1$: Beginning of the Simulation

Below is an example after the vertical line has been changed. The distances to the line have been recalculated.

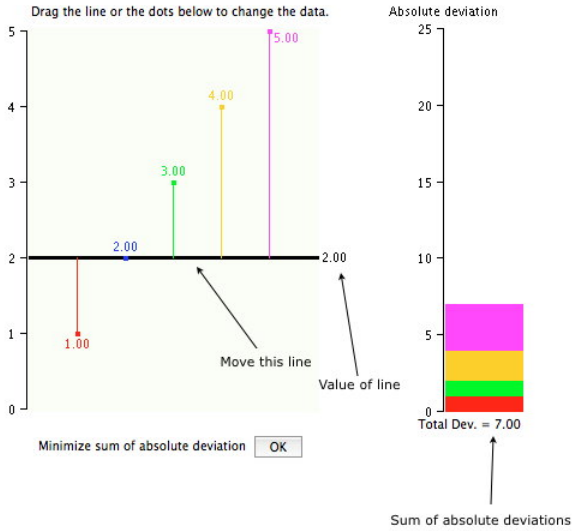


Figure 3.5.2: *Simulation after change*

You can change the data by clicking on a data point and dragging. An example with changed data is shown below.

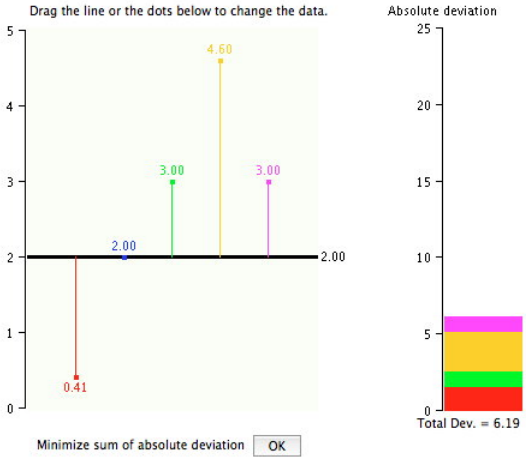


Figure 3.5.3: *Simulation after changing data point*

3.6: Squared Differences Simulation

Learning Objectives

- Learn which measure of central tendency minimizes the sum of squared deviations

Instructions

This demonstration allows you to examine the sum of squared deviations from a given value. The graph to the right shows the numbers $1, 2, 3, 4,$ and 5 and their deviations from an arbitrary starting value of 0.254 (the figure displays this rounded to 0.25).

The first number, 1, is represented by a red dot. The deviation from 0.254 is represented by a red line from the red dot to the black line. The value of the black line is 0.254.

Similarly, the number 2 is represented by a blue dot and its deviation from 0.25 is represented by a blue line.

The height of the colored rectangles represents the sum of the absolute deviations from the black line. The sum of the deviations of the numbers $1, 2, 3, 4,$ and 5 from 0.25 is $0.746 + 1.75 + 2.746 + 3.746 + 4.746 = 13.73$.

The area of each rectangle represents the magnitude of the squared deviation of a point from the black line. For example, the red rectangle has an area of $0.746 \times 0.746 = 0.557$. The sum of all the areas of the rectangles is 47.70. The sum of all the areas represents the sum of the squared deviations.

In this demonstration, you can move the black bar by clicking on it and dragging it up or down. In the second figure it has been moved up to 1.0. The deviation of the red point from the black bar is now 0 since they are both 1. The sum of the deviations is now $10(0+1+2+3+4)$ and the sum of squared deviations is $30(0^2+1^2+2^2+3^2+4^2)$.

As you move the bar up and down, the value of the sum of absolute deviations and the sum of squared deviations changes.

See if you can find the placement of the black bar that produces the smallest value for the sum of the squared deviations.

To check and see if you found the smallest value, click the "OK" button at the bottom of the graph. It will move the bar to the location that produces the smallest sum of squared deviations.

For the initial data, the value that minimizes the sum of squared deviations is also the value that minimizes the sum of absolute deviations. This will not be true for most data.

You can also move the individual points. Click on one of the points and move it up or down and note the effect. Your goal for this demonstration is to discover a rule for determining what value will give you the smallest sum of squared deviations.

Illustrated Instructions

Below is a screen shot of the simulation's beginning screen. The "Total Dev." is the sum of the absolute deviations; the "Total Area" is the sum of the squared deviations.

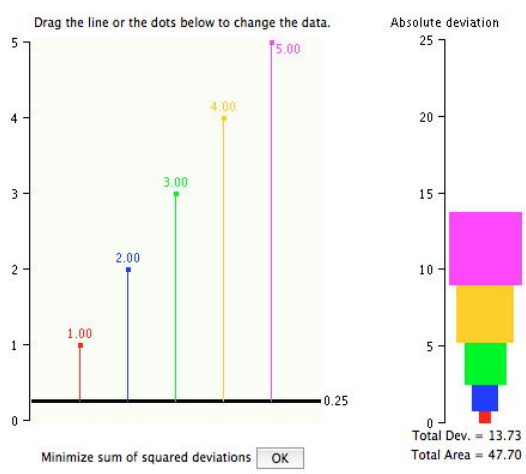


Figure 3.6.1: Beginning Screen

Below is an example after the vertical line has been changed. The distances to the line have been recalculated.

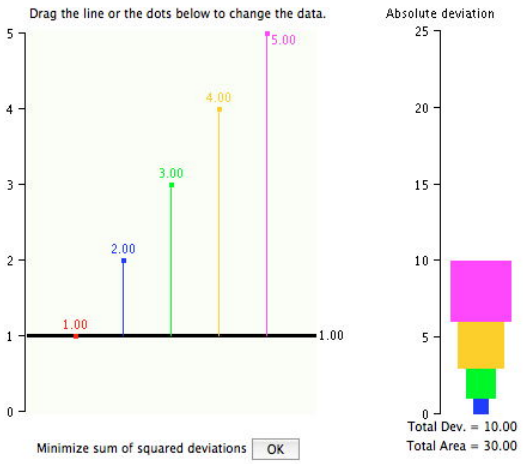


Figure 3.6.2: Screen shot after change

You can change the data by clicking on a data point and dragging. An example with changed data is shown below.

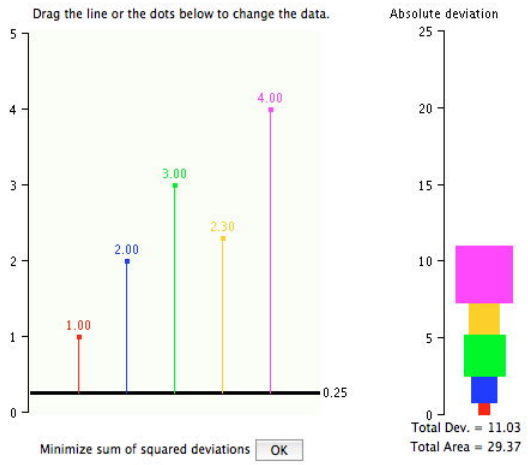


Figure 3.6.3: Screen *shot after changing the data*

3.7: Median and Mean

Learning Objectives

- State whether it is the mean or median that minimizes the mean absolute deviation
- State whether it is the mean or median that is the balance point on a balance scale

In the section "What is central tendency," we saw that the center of a distribution could be defined three ways:

1. the point on which a distribution would balance
2. the value whose average absolute deviation from all the other values is minimized
3. the value whose average squared difference from all the other values is minimized

From the simulation in this chapter, you discovered (we hope) that the mean is the point on which a distribution would balance, the median is the value that minimizes the sum of absolute deviations, and the mean is the value that minimizes the sum of the squared deviations.

Table $3.7.1$ shows the absolute and squared deviations of the numbers $2, 3, 4, 9$ and 16 from their median of 4 and their mean of 6.8. You can see that the sum of absolute deviations from the median (20) is smaller than the sum of absolute deviations from the mean (22.8). On the other hand, the sum of squared deviations from the median (174) is larger than the sum of squared deviations from the mean (134.8).

Table $3.7.1$: Absolute and squared deviations from the median of 4 and the mean of 6.8

Value	Absolute Deviation from Median	Absolute Deviation from Mean	Squared Deviation from Median	Squared Deviation from Mean
2	2	4.8	4	23.04
3	1	3.8	1	14.44
4	0	2.8	0	7.84
9	5	2.2	25	4.84
16	12	9.2	144	84.64
Total	20	22.8	174	134.8

Figure $3.7.1$ shows that the distribution balances at the mean of 6.8 and not at the median of 4. The relative advantages and disadvantages of the mean and median are discussed in the section "Comparing Measures" later in this chapter.

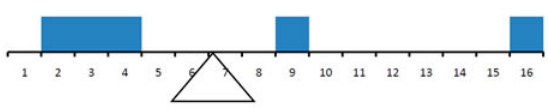


Figure $3.7.1$: The distribution balances at the mean of 6.8 and not at the median of 4.0.

When a distribution is symmetric, then the mean and the median are the same. Consider the following distribution: $1, 3, 4, 5, 6, 7, 9$ The mean and median are both 5. The mean, median, and mode are identical in the bell-shaped normal distribution.

3.8: Mean and Median Demo

Learning Objectives

- To study how the mean and median change with different distributions

Instructions

This demonstration shows how the relative size of the mean and the median depends on the skew of the distribution. The demonstration begins by showing a histogram of a symmetric distribution (no skew). The mean and median are both 5.0. The mean is shown on the histogram as a small blue line; the median is shown as a small purple line. The standard deviation is 1.81. A red line extends one sd in each direction from the mean. The standard deviation is calculated assuming the data portrayed in the graph represent the entire population. You can change the values of the data set by "painting" the histogram with the mouse. Change the distribution in various ways and note how the skew affects whether the mean is bigger than the median or vice versa.

Illustrated Instructions

The simulation starts out with a symmetrical distribution. As can be seen seen in the screenshot below the mean and median are equal and there is 0 skew.

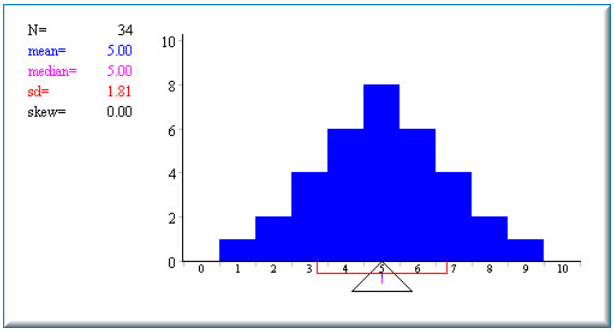


Figure $3.8.1$: Beginning of the Simulation

The distribution can be changed "painting" it with the mouse. Below is an example of a negatively skewed distribution. Note that the mean and median are no longer equal to each other. Trying painting several different types of distributions to see how mean and median values are affected relative to each other.

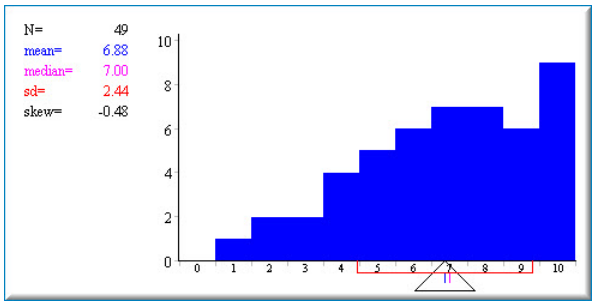


Figure $3.8.2$: Negatively Skewed Distribution

3.9: Additional Measures

Learning Objectives

- Compute the trimean
- Compute the geometric mean directly
- Compute the geometric mean using logs
- Use the geometric to compute annual portfolio returns
- Compute a trimmed mean

Although the mean, median, and mode are by far the most commonly used measures of central tendency, they are by no means the only measures. This section defines three additional measures of central tendency: the trimean, the geometric mean, and the trimmed mean. These measures will be discussed again in the section "Comparing Measures of Central Tendency."

Trimean

The trimean is a weighted average of the 25th percentile, the 50th percentile, and the 75th percentile. Letting $P25$ be the 25th percentile, $P50$ be the 50th and $P75$ be the 75th percentile, the formula for the trimean is:

$$Trimean = \frac{P25 + 2P50 + P75}{4} \tag{3.9.1}$$

As you can see from the formula, the median is weighted twice as much as the 25th and 75th percentiles. Table $3.9.1$ shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The relevant percentiles are shown in Table $3.9.2$.

Table $3.9.1$: *Number of touchdown passes*

$$\begin{array}{cccccccc} 37 & 33 & 33 & 32 & 29 & 28 & 28 & 23 \\ 22 & 22 & 22 & 21 & 21 & 21 & 20 & 20 \\ 19 & 19 & 18 & 18 & 18 & 18 & 16 & 15 \\ 14 & 14 & 14 & 12 & 12 & 9 & 6 & \end{array} \tag{3.9.2}$$

Table $3.9.2$: *Percentiles*

Percentile	Value
25	15
50	20
75	23

The trimean is therefore

$$\frac{15 + 2 \times 20 + 23}{4} = \frac{78}{4} = 19.5. \tag{3.9.3}$$

Geometric Mean

The geometric mean is computed by multiplying all the numbers together and then taking the nth root of the product. For example, for the numbers $1, 10$ and 100, the product of all the numbers is:

$$1 \times 10 \times 100 = 1,000. \tag{3.9.4}$$

Since there are three numbers, we take the cubed root of the product $(1,000)$ which is equal to 10. The formula for the geometric mean is therefore

$$(\Pi X)^{1/N} \tag{3.9.5}$$

where the symbol Π means to multiply. Therefore, the equation says to multiply all the values of X and then raise the result to the $1/N$th power. Raising a value to the $\frac{1}{N}^{th}$ power is, of course, the same as taking the Nth root of the value. In this case, $1000^{1/3}$ is the cube root of $1,000$.

The geometric mean has a close relationship with logarithms. Table $3.9.3$ shows the logs (base 10) of these three numbers. The arithmetic mean of the three logs is 1. The anti-log of this **arithmetic mean** of 1 is the **geometric mean**. The anti-log of 1 is $10^1 = 10$. Note that the geometric mean only makes sense if all the numbers are positive.

Table $3.9.3$: *Logarithms*

X	$\log_{10}(X)$
1	0
10	1
100	2

The geometric mean is an appropriate measure to use for averaging rates. For example, consider a stock portfolio that began with a value of $\$1,000$ and had annual returns of $13\%, 22\%, 12\%, -5\%, and -13\%$ Table $3.9.4$ shows the value after each of the five years.

Table $3.9.4$: *Portfolio Returns*

Year	Return	Value
1	13%	1,130
2	22%	1,379
3	12%	1,544
4	-5%	1,467
5	-13%	1,276

The question is how to compute average annual rate of return. The answer is to compute the geometric mean of the returns. Instead of using the percents, each return is represented as a multiplier indicating how much higher the value is after the year. This multiplier is 1.13 for a 13% return and 0.95 for a 5% loss. The multipliers for this example are $1.13, 1.22, 1.12, 0.95, and\ 0.87$ The geometric mean of these multipliers is 1.05. Therefore, the average annual rate of return is 5%. Table $3.9.5$ shows how a portfolio gaining 5% a year would end up with the same value ($\$1,276$) as shown in Table $3.9.4$.

Table $3.9.5$: *Portfolio Returns*

Year	Return	Value
1	5%	1,050
2	5%	1,103
3	5%	1,158
4	5%	1,216
5	5%	1,276

Trimmed Mean

To compute a trimmed mean, you remove some of the higher and lower scores and compute the mean of the remaining scores. A mean trimmed 10% is a mean computed with 10% of the scores trimmed off: 5% from the bottom and 5% from the top. A mean trimmed 50% is computed by trimming the upper 25% of the scores and the lower 25% of the scores and computing the mean of

the remaining scores. The trimmed mean is similar to the median which, in essence, trims the upper 49% and the lower 49% of the scores. Therefore the trimmed mean is a hybrid of the mean and the median. To compute the mean trimmed 20% for the touchdown pass data shown in Table $3.9.1$, you remove the lower 10% of the scores $(6, 9, \ and \ 12)$ as well as the upper 10% of the scores $(33, 33, \ and \ 37)$ and compute the mean of the remaining 25 scores. This mean is 20.16.

3.10: Comparing Measures

Learning Objectives

- State how the measures differ in symmetric distributions
- State which measure(s) should be used to describe the center of a skewed distribution

How do the various measures of central tendency compare with each other? For symmetric distributions, the mean, median, trimean, and trimmed mean are equal, as is the mode except in bimodal distributions. Differences among the measures occur with skewed distributions. Figure 3.10.1 shows the distribution of 642 scores on an introductory psychology test. Notice this distribution has a slight positive skew.

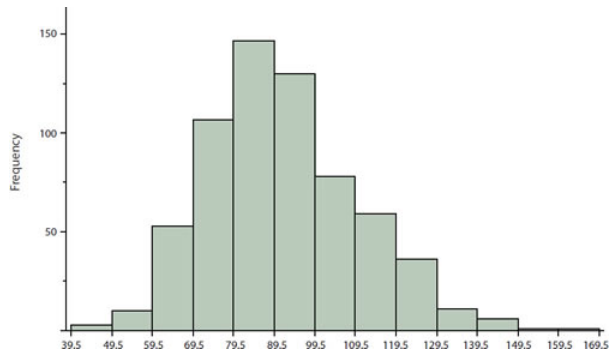


Figure 3.10.1: *A distribution with a positive skew.*

Measures of central tendency are shown in Table 3.10.1. Notice they do not differ greatly, with the exception that the mode is considerably lower than the other measures. When distributions have a positive skew, the mean is typically higher than the median, although it may not be in bimodal distributions. For these data, the mean of 91.58 is higher than the median of 90. Typically the trimean and trimmed mean will fall between the median and the mean, although in this case, the trimmed mean is slightly lower than the median. The geometric mean is lower than all measures except the mode.

Table 3.10.1: *Measures of central tendency for the test scores*

Measure	Value
Mode	84.00
Median	90.00
Geometric Mean	89.70
Trimean	90.25
Mean Trimmed 50%	89.81
Mean	91.58

The distribution of baseball salaries (in 1994) shown in Figure 3.10.2 has a much more pronounced skew than the distribution in Figure 3.10.2

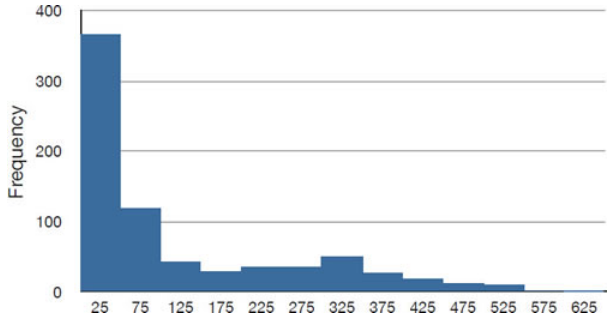


Figure 3.10.2: A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars: 25 equals $250,000$).

Table $3.10.2$ shows the measures of central tendency for these data. The large skew results in very different values for these measures. No single measure of central tendency is sufficient for data such as these. If you were asked the very general question: "So, what do baseball players make?" and answered with the mean of $\$1,183,000$ you would not have told the whole story since only about one third of baseball players make that much. If you answered with the mode of $\$250,000$ or the median of $\$500,000$ you would not be giving any indication that some players make many millions of dollars. Fortunately, there is no need to summarize a distribution with a single number. When the various measures differ, our opinion is that you should report the mean, median, and either the trimean or the mean trimmed 50%. Sometimes it is worth reporting the mode as well. In the media, the median is usually reported to summarize the center of skewed distributions. You will hear about median salaries and median prices of houses sold, etc. This is better than reporting only the mean, but it would be informative to hear more statistics.

Table $3.10.2$: Measures of central tendency for baseball salaries (in thousands of dollars)

Measure	Value
Mode	250
Median	500
Geometric Mean	555
Trimean	792
Mean Trimmed 50%	619
Mean	1,183

3.11: Variability

Learning Objectives

- To study how much the numbers in a distribution differ from each other

Variability refers to how much the numbers in a distribution differ from each other. The most common measures are presented in "Measures of Variability." The "variability demo" allows you to change the standard deviation of a distribution and view a graph of the changed distribution.

One of the more counter-intuitive facts in introductory statistics is that the formula for variance when computed in a population is biased when applied in a sample. The "Estimating Variance Simulation" shows concretely why this is the case.

3.12: Measures of Variability

Learning Objectives

- Compute the range
- Compute the variance in the population
- Compute the standard deviation from the variance

What is Variability?

Variability refers to how "spread out" a group of scores is. To see what we mean by spread out, consider graphs in Figure $3.12.1$. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.

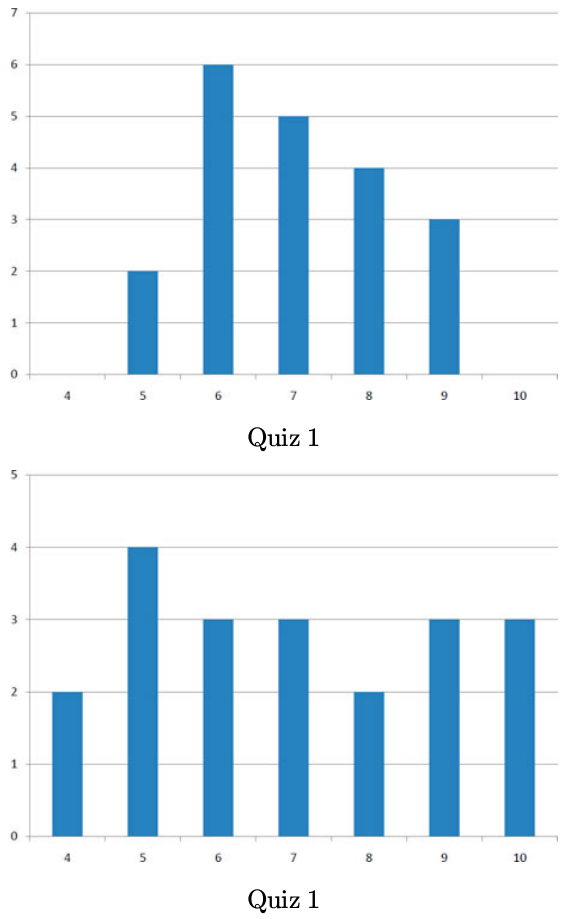


Figure $3.12.1$: *Bar charts of two quizzes*

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the center of a distribution of scores, in this chapter we will discuss measures of the variability of a distribution. There are four frequently used measures of variability: the range, interquartile range, variance, and standard deviation. In the next few paragraphs, we will look at each of these four measures of variability in more detail.

Range

The range is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score. Let's take a few examples. What is the range of the following group of numbers: $10, 2, 5, 6, 7, 3, 4$? Well, the highest number is 10, and the lowest number is 2, so $10 - 2 = 8$. The range is 8. Let's take another example. Here's a dataset with 10 numbers: $99, 45, 23, 67, 45, 91, 82, 78, 62, 51$. What is the range? The highest number is

99 and the lowest number is 23, so $99-23$ equals 76; the range is 76. Now consider the two quizzes shown in Figure 3.12.1. On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4. The range on Quiz 2 was larger: the lowest score was 4 and the highest score was 10. Therefore the range is 6.

Interquartile Range

The interquartile range (IQR) is the range of the middle 50% of the scores in a distribution. It is computed as follows:

$$IQR = 75^{th} percentile - 25^{th} percentile \tag{3.12.1}$$

For Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. For Quiz 2, which has greater spread, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4. Recall that in the discussion of box plots, the 75th percentile was called the upper hinge and the 25th percentile was called the lower hinge. Using this terminology, the interquartile range is referred to as the H-spread.

A related measure of variability is called the semi-interquartile range. The semi-interquartile range is defined simply as the interquartile range divided by 2. If a distribution is symmetric, the median plus or minus the semi-interquartile range contains half the scores in the distribution.

Variance

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, the variance is defined as the average squared difference of the scores from the mean. The data from Quiz 1 are shown in Table 3.12.1. The mean score is 7.0. Therefore, the column "Deviation from Mean" contains the score minus 7. The column "Squared Deviation" is simply the previous column squared.

Table 3.12.1: Calculation of Variance for Quiz 1 scores

Scores	Deviation from Mean	Squared Deviation
9	2	4
9	2	4
9	2	4
8	1	1
8	1	1
8	1	1
8	1	1
7	0	0
7	0	0
7	0	0
7	0	0
7	0	0
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1

6	-1	1
5	-2	4
5	-2	4
Means		
7	0	1.5

One thing that is important to notice is that the mean deviation from the mean is 0. This will always be the case. The mean of the squared deviations is 1.5. Therefore, the variance is 1.5. Analogous calculations with $\text{Quiz } 2$ show that its variance is 6.7. The formula for the variance is:

$$s^2 = \frac{\sum(X-\mu)^2}{N} \tag{3.12.2}$$

where σ^2 is the variance, μ is the mean, and N is the number of numbers. For $\text{Quiz } 1$, $\mu = 7$ and $N = 20$.

If the variance in a sample is used to estimate the variance in a population, then the previous formula underestimates the variance and the following formula should be used:

$$s^2 = \frac{\sum(X-M)^2}{N-1} \tag{3.12.3}$$

where s^2 is the estimate of the variance and M is the sample mean. Note that M is the mean of a sample taken from a population with a mean of μ. Since, in practice, the variance is usually computed in a sample, this formula is most often used. The simulation "estimating variance" illustrates the bias in the formula with N in the denominator.

Let's take a concrete example. Assume the scores $1, 2, 4, \ and \ 5$ were sampled from a larger population. To estimate the variance in the population you would compute s^2 as follows:

$$M = \frac{1+2+4+5}{4} = \frac{12}{4} = 3 \tag{3.12.4}$$

$$\begin{aligned} s^2 &= \frac{[(1-3)^2+(2-3)^2+(4-3)^2+(5-3)^2]}{(4-1)} \\ &= \frac{(4+1+1+4)}{3} \\ &= \frac{10}{3} \\ &= 3.333 \end{aligned}$$

There are alternate formulas that can be easier to use if you are doing your calculations with a hand calculator. You should note that these formulas are subject to rounding error if your values are very large and/or you have an extremely large number of observations.

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N} \tag{3.12.5}$$

and

$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1} \tag{3.12.6}$$

For this example,

$$\sum X^2 = 1^2 + 2^2 + 4^2 + 5^2 = 46 \tag{3.12.7}$$

$$\frac{(\sum X)^2}{N} = \frac{(1+2+4+5)^2}{4} = \frac{144}{4} = 36 \tag{3.12.8}$$

$$\sigma^2 = \frac{(46-36)}{4} = 2.5 \tag{3.12.9}$$

$$s^2 = \frac{(46-36)}{3} = 3.333 \text{ as with the other formula} \tag{3.12.10}$$

Standard Deviation

The standard deviation is simply the square root of the variance. This makes the standard deviations of the two quiz distributions 1.225 and 2.588. The standard deviation is an especially useful measure of variability when the distribution is normal or approximately normal (see Chapter on Normal Distributions) because the proportion of the distribution within a given number of standard deviations from the mean can be calculated. For example, 68% of the distribution is within one standard deviation of the mean and approximately 95% of the distribution is within two standard deviations of the mean. Therefore, if you had a normal distribution with a mean of 50 and a standard deviation of 10, then 68% of the distribution would be between $50-10=40$ and $50+10=60$. Similarly, about 95% of the distribution would be between $50-2\times10=30$ and $50+2\times10=70$. The symbol for the population standard deviation is σ; the symbol for an estimate computed in a sample is s. Figure $3.12.2$ shows two normal distributions. The red distribution has a mean of 40 and a standard deviation of 5; the blue distribution has a mean of 60 and a standard deviation of 10. For the red distribution, 68% of the distribution is between 35 and 45; for the blue distribution, 68% is between 50 and 70.

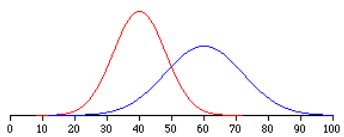


Figure $3.12.2$: *Normal distributions with standard deviations of 5 and 10*

3.13: Variability Demo

Learning Objectives

- Identify differences in the means and standard deviations of distributions

Instructions

The demonstration shows a graph of two normal distributions. Both distributions have means of 50. The red distribution has a standard deviation of 10; the blue distribution has a standard deviation of 5. You can see that the red distribution is more spread out than the blue distribution. Note that about two thirds of the area of the distributions is within one standard deviation of the mean. For the red distribution, this is between 40 and 60; for the blue distribution, this is between 45 and 55. About 95% of a normal distribution is within two standard deviations from the mean. For the red distribution, this is between 30 and 70; for the blue it is between 40 and 60.

You can change the means and standard deviations of the distributions and see the results visually. For some values, the distributions will be off the graph. For example, if you give a distribution a mean of 200, it will not be shown.

Illustrated Instructions

The demonstration starts with 2 normal distributions with equal means and different standard deviations (see screenshot below).

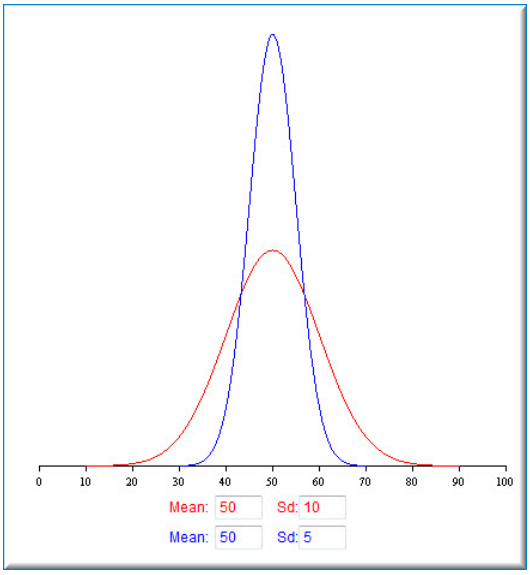


Figure 3.13.1*: Normal Distributions variability demo*

The means and standard deviations for both distributions can be changed and these changes will be reflected in the graph. The screenshot below shows the distributions with different means and standard deviations.

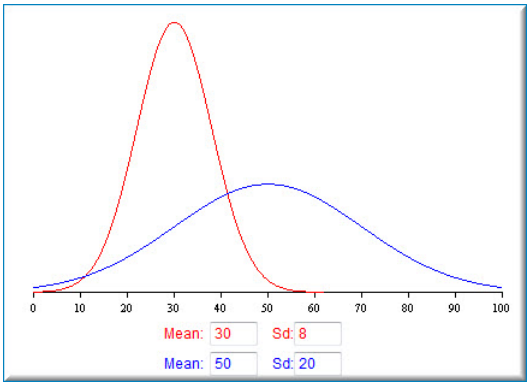


Figure 3.13.2*: D*istributions with different means and standard deviations

3.14: Estimating Variance Simulation

Learning Objectives

- Learn which measure of central tendency will balance a distribution

Instructions

This simulation samples from the population of 50 numbers shown here. You can see that there are 10 instances of the values $1, 2, 3, 4$, and 5. The mean of the population is therefore 3. The variance is the average squared deviation from the mean of 3. You can compute that this is exactly 2.

When you click on the button "Draw 4 numbers" four scores are sampled (with replacement) from the population. The four numbers are shown in red, as is the mean of the four numbers. The variance is then computed in two ways. The upper formula computes the variance by computing the mean of the squared deviations or the four sampled numbers from the sample mean. The lower formula computes the mean of the squared deviations or the four sampled numbers from the population mean of 3.00 (on rare occasions, the sample and population means will be equal). The computed variances are placed in the fields to the right of the formulas. The mean of the values in a field is shown at the bottom of the field. When there is only one value in the field, the mean will, of course, equal that value.

If you click the "Draw 4 numbers" button again, another four numbers will be sampled. The mean and variance will also be computed as before. The fields to the right of the formulas will hold both variances and the bottom of the field will show the mean of the variances.

The population variance is exactly 2. Use this fact to assess the relative value of the two formulas for variance. See which one, on average, approaches 2 and which one gives lower estimates. Explore whether either formula is always more accurate, or whether sometimes one is more accurate and at other times, the other formula is. If the variance based on the sample mean had been computed by dividing by $N-1=3$ instead of 4, then the variance would be $\frac{4}{3}$ times bigger. Does multiplying the variance by $\frac{4}{3}$ lead to better estimates?

Illustrated Instructions

As can be seen in the screenshot below, the variance estimation simulation begins by displaying a population of 50 numbers ranging from $1-5$.

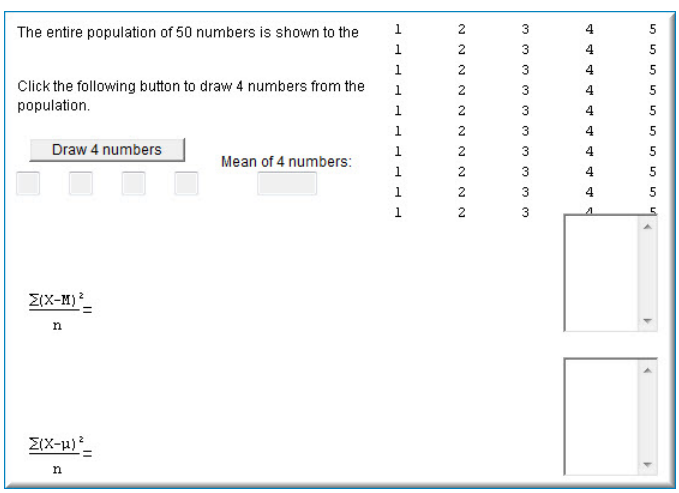


Figure 3.14.1: Beginning Variance Estimation Simulation

Each time the "Draw 4 numbers" button is clicked four numbers are sampled from the population and the mean, the variance of the sample from the sample mean as well as the variance of the sample from the population mean are calculated. The variances are stored in fields next to their respective formula. The screenshot below shows the simulation after the "Draw 4 numbers" button has been clicked four times.

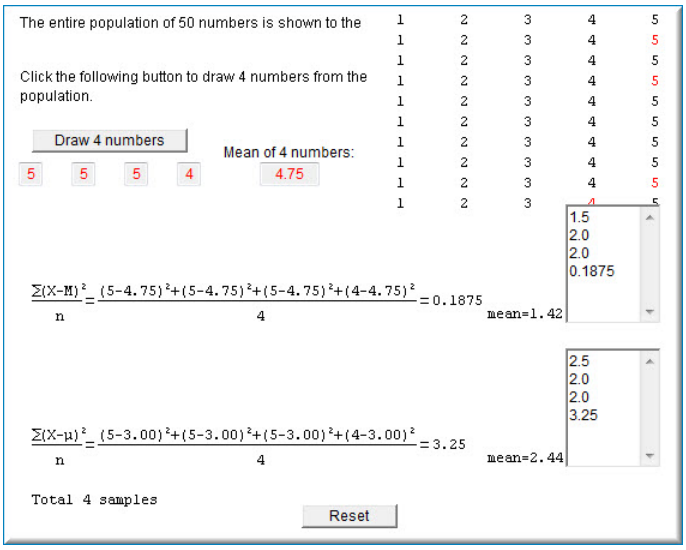


Figure 3.14.2: Simulation after "Draw 4 numbers" button is clicked

Use the simulation to explore whether either formula is on average more accurate than the other.

3.15: Shapes of Distributions

Learning Objectives

- Compute skew using two different formulas
- Compute kurtosis

We saw in the section on distributions in Chapter 1 that shapes of distributions can differ in skew and/or kurtosis. This section presents numerical indexes of these two measures of shape.

Skew

Figure $3.15.1$ shows a distribution with a very large positive skew. Recall that distributions with positive skew have tails that extend to the right.

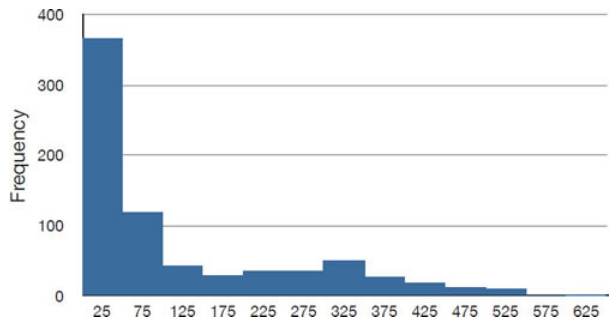


Figure 3.15.1: A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in tens of thousands of dollars).

Distributions with positive skew normally have larger means than medians. The mean and median of the baseball salaries shown in Figure $3.15.1$ are $\$1,183,417$and $\$500,000$respectively. Thus, for this highly-skewed distribution, the mean is more than twice as high as the median. The relationship between skew and the relative size of the mean and median led the statistician Pearson to propose the following simple and convenient numerical index of skew:

$$\frac{3(Mean - Median)}{\sigma} \tag{3.15.1}$$

The standard deviation of the baseball salaries is $1,390,922$ Therefore, Pearson's measure of skew for this distribution is $\frac{3(1,183,417 - 500,000)}{1,390,922} = 1.47.$

Just as there are several measures of central tendency, there is more than one measure of skew. Although Pearson's measure is a good one, the following measure is more commonly used. It is sometimes referred to as the third moment about the mean.

$$\sum \frac{(X-\mu)^3}{\sigma^3} \tag{3.15.2}$$

Kurtosis

The following measure of kurtosis is similar to the definition of skew. The value "3" is subtracted to define "no kurtosis" as the kurtosis of a normal distribution. Otherwise, a normal distribution would have a kurtosis of 3.

$$\sum \frac{(X-\mu)^4}{\sigma^4} - 3 \tag{3.15.3}$$

3.16: Comparing Distributions Demo

Learning Objectives

- Learn how one might compare data from two distributions

Instructions

This demonstration collects response time data from you and computes various descriptive statistics. It also displays histograms and box plots. After you click the button to begin, a rectangle will appear to the right of the button. Move the cursor to the rectangle and click. The time it takes you to click on the rectangle is recorded. Half of the time the rectangle is small and half of the time it is big. After two practice trials and 40 regular trials, the descriptive statistics and graphs will be displayed.

Illustrated Instructions

The simulation is started by clicking the button in the lower left of the screen. Clicking this button begins a series of trials where a another button appears on the right and you will be asked to click this button as quickly as possible.

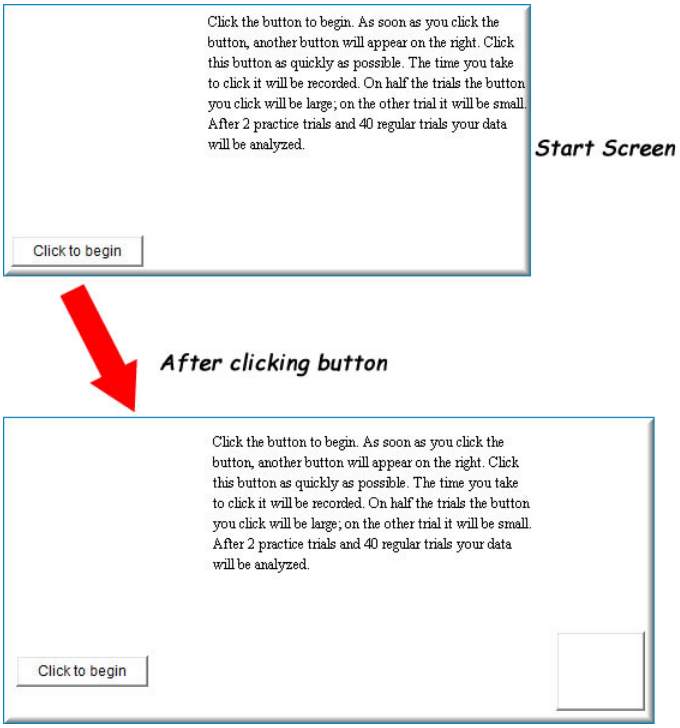
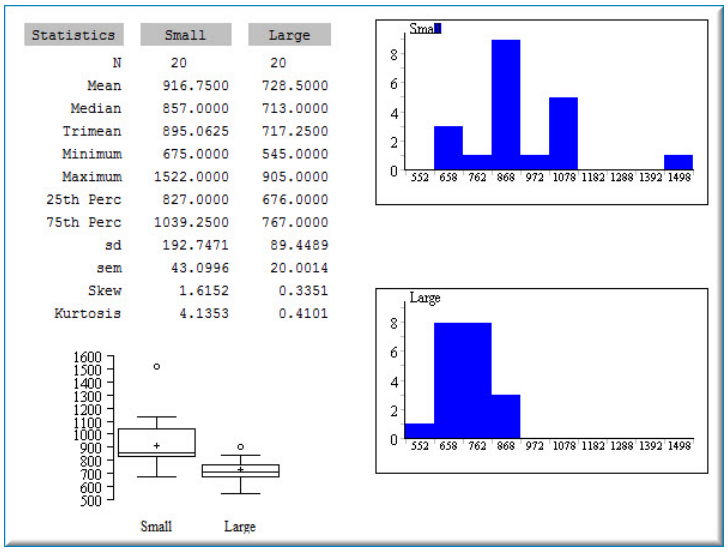


Figure 3.16.1: Distributions Simulation

Half of the buttons will large and half will be small. Once you complete 40 trials, the simulation will show you descriptive statistics about the reaction time data for each button size (see screenshot below).



Statistics	Small	Large
N	20	20
Mean	916.7500	728.5000
Median	857.0000	713.0000
Trimean	895.0625	717.2500
Minimum	675.0000	545.0000
Maximum	1522.0000	905.0000
25th Perc	827.0000	676.0000
75th Perc	1039.2500	767.0000
sd	192.7471	89.4489
sem	43.0996	20.0014
Skew	1.6152	0.3351
Kurtosis	4.1353	0.4101

Figure 3.16.2: Simulation showing descriptive statistics after 40 trials

3.17: Effects of Linear Transformations

Learning Objectives

- Compute the mean of a transformed variable
- Compute the variance of a transformed variable

This section covers the effects of linear transformations on measures of central tendency and variability. Let's start with an example we saw before in the section that defined linear transformation: temperatures of cities. Table $3.17.1$ shows the temperatures of 5 cities.

Table $3.17.1$: Temperatures in 5 cities on $11/16/2002$

City	Degrees Fahrenheit	Degrees Centigrade
Houston	54	12.22
Chicago	37	2.78
Minneapolis	31	-0.56
Miami	78	25.56
Phoenix	70	21.11
Mean	54.000	12.220
Median	54.000	12.220
Variance	330.00	101.852
SD	18.166	10.092

Recall that to transform the degrees Fahrenheit to degrees Centigrade, we use the formula

$$C = 0.556F - 17.778 \tag{3.17.1}$$

which means we multiply each temperature Fahrenheit by 0.556 and then subtract 17.778. As you might have expected, you multiply the mean temperature in Fahrenheit by 0.556 and then subtract 17.778 to get the mean in Centigrade. That is, $(0.556)(54) - 17.778 = 12.22$ The same is true for the median. Note that this relationship holds even if the mean and median are not identical as they are in Table $3.17.1$.

The formula for the standard deviation is just as simple: the standard deviation in degrees Centigrade is equal to the standard deviation in degrees Fahrenheit times 0.556. Since the variance is the standard deviation squared, the variance in degrees Centigrade is equal to 0.5562^2 times the variance in degrees Fahrenheit.

To sum up, if a variable X has a mean of μ, a standard deviation of σ, and a variance of σ^2, then a new variable Y created using the linear transformation

$$Y = bX + A \tag{3.17.2}$$

will have a mean of $b\mu + A$, a standard deviation of $b\sigma$, and a variance of $b^2\sigma^2$.

It should be noted that the term "linear transformation" is defined differently in the field of linear algebra. For details, follow this link.

3.18: Statistical Literacy

Learning Objectives

- How to select between Mean and Median

The Mean or the Median?

Example 3.18.1

The playbill for the Alley Theatre in Houston wants to appeal to advertisers. They reported the mean household income and the median age of theatergoers. What might have guided their choice of the mean or median?

Solution

It is likely that they wanted to emphasize that theatergoers had high income but de-emphasize how old they are. The distributions of income and age of theatergoers probably have positive skew. Therefore the mean is probably higher than the median, which results in higher income and lower age than if the median household income and mean age had been presented.

3.E: Summarizing Distributions (Exercises)

General questions

Q1

Make up a dataset of 12 numbers with a positive skew. Use a statistical program to compute the skew. Is the mean larger than the median as it usually is for distributions with a positive skew? What is the value for skew? (relevant section & relevant section)

Q2

Repeat **Q1** only this time make the dataset have a negative skew. (relevant section & relevant section)

Q3

Make up three data sets with 5 numbers each that have: (relevant section & relevant section)

a. the same mean but different standard deviations.
b. the same mean but different medians.
c. the same median but different means.

Q4

Find the mean and median for the following three variables: (relevant section)

$$\begin{array}{ccc} A & B & C \\ 8 & 4 & 6 \\ 5 & 4 & 2 \\ 7 & 6 & 3 \\ 1 & 3 & 4 \\ 3 & 4 & 1 \end{array} \tag{3.E.1}$$

Q5

A sample of 30 distance scores measured in yards has a mean of 7, a variance of 16, and a standard deviation of 4.

a. You want to convert all your distances from yards to feet, so you multiply each score in the sample by 3. What are the new mean, variance, and standard deviation?
b. You then decide that you only want to look at the distance past a certain point. Thus, after multiplying the original scores by 3, you decide to subtract 4 feet from each of the scores. Now what are the new mean, variance, and standard deviation? (relevant section)

Q6

You recorded the time in seconds it took for 8 participants to solve a puzzle. These times appear below. However, when the data was entered into the statistical program, the score that was supposed to be 22.1 was entered as 21.2. You had calculated the following measures of central tendency: the mean, the median, and the mean trimmed 25%. Which of these measures of central tendency will change when you correct the recording error? (relevant section & relevant section)

$$\begin{array}{c} 15.2 \\ 18.8 \\ 19.3 \\ 19.7 \\ 20.2 \\ 21.8 \\ 22.1 \\ 29.4 \end{array} \tag{3.E.2}$$

Q7

For the test scores in question **Q6**, which measures of variability (range, standard deviation, variance) would be changed if the 22.1 data point had been erroneously recorded as 21.2? (relevant section)

Q8

You know the minimum, the maximum, and the 25th, 50th, and 75th percentiles of a distribution. Which of the following measures of central tendency or variability can you determine? (relevant section, relevant section & relevant section)

mean, median, mode, trimean, geometric mean,

range, interquartile range, variance, standard deviation

Q9

a. Find the value (v) for which $\sum(X-v)^2$ is minimized.
b. Find the value (v) for which $\sum|X-v|$ is minimized.

Q10

Your younger brother comes home one day after taking a science test. He says that someone at school told him that "60% of the students in the class scored above the median test grade." What is wrong with this statement? What if he said "60% of the students scored below the mean?" (relevant section)

Q11

An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess positions. Compare the performance of each group. Consider spread as well as central tendency. (relevant section, relevant section & relevant section)

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

Q12

True/False: A bimodal distribution has two modes and two medians. (relevant section)

Q13

True/False: The best way to describe a skewed distribution is to report the mean. (relevant section)

Q14

True/False: When plotted on the same graph, a distribution with a mean of 50 and a standard deviation of 10 will look more spread out than will a distribution with a mean of 60 and a standard deviation of 5. (relevant section)

Q15

Compare the mean, median, trimean in terms of their sensitivity to extreme scores (relevant section).

Q16

If the mean time to respond to a stimulus is much higher than the median time to respond, what can you say about the shape of the distribution of response times? (relevant section)

Q17

A set of numbers is transformed by taking the log base 10 of each number. The mean of the transformed data is 1.65. What is the geometric mean of the untransformed data? (relevant section)

Q18

Which measure of central tendency is most often used for returns on investment?

Q19

The histogram is in balance on the fulcrum. What are the mean, median, and mode of the distribution (approximate where necessary)?

Questions from Case Studies

The following questions are from the Angry Moods (AM) case study.

Q20

(AM#4) Does Anger-Out have a positive skew, a negative skew, or no skew? (relevant section)

Q21

(AM#8) What is the range of the Anger-In scores? What is the interquartile range? (relevant section)

Q22

(AM#12) What is the overall mean Control-Out score? What is the mean Control-Out score for the athletes? What is the mean Control-Out score for the non-athletes? (relevant section)

Q23

(AM#15) What is the variance of the Control-In scores for the athletes? What is the variance of the Control-In scores for the non-athletes? (relevant section)

The following question is from the Flatulence (F) case study.

Q24

(F#2) Based on a histogram of the variable "perday", do you think the mean or median of this variable is larger? Calculate the mean and median to see if you are right. (relevant section & relevant section)

The following questions are from the Stroop (S) case study.

Q25

(S#1) Compute the mean for "words". (relevant section)

Q26

(S#2) Compute the mean and standard deviation for "colors". (relevant section & relevant section)

The following questions are from the Physicians' Reactions (PR) case study.

Q27

(PR#2) What is the mean expected time spent for the average-weight patients? What is the mean expected time spent for the overweight patients? (relevant section)

Q28

(PR#3) What is the difference in means between the groups? By approximately how many standard deviations do the means differ? (relevant section & relevant section)

The following question is from the Smiles and Leniency (SL) case study.

Q29

(SL#2) Find the mean, median, standard deviation, and interquartile range for the leniency scores of each of the four groups. (relevant section & relevant section)

The following questions are from the ADHD Treatment (AT) case study.

Q30

(AT#4) What is the mean number of correct responses of the participants after taking the placebo (0 mg/kg)? (relevant section)

Q31

(AT#7) What are the standard deviation and the interquartile range of the $d0$ condition? (relevant section)

Selected Answers

S4

Variable A: Mean = 4.8, Median = 5

S5

a. Mean = 21, Var = 144, SD = 12

S9

a. 5.2

S22

Non-athletes: 23.2

S23

Athletes: 20.5

S26

Mean = 20.2

S27

Ave. weight: 31.4

S29

False smile group:

Mean = 5.37

Median = 5.50

SD = 1.83

IQR = 3.0

CHAPTER OVERVIEW

4: Normal Distribution

Most of the statistical analyses presented in this book are based on the bell-shaped or normal distribution. The introductory section defines what it means for a distribution to be normal and presents some important properties of normal distributions. The interesting history of the discovery of the normal distribution is described in the second section. Methods for calculating probabilities based on the normal distribution are described in Areas of Normal Distributions. The Varieties of Normal Distribution Demo allows you to enter values for the mean and standard deviation of a normal distribution and see a graph of the resulting distribution. A frequently used normal distribution is called the Standard Normal distribution and is described in the section with that name. The binomial distribution can be approximated by a normal distribution. The section Normal Approximation to the Binomial shows this approximation. The Normal Approximation Demo allows you to explore the accuracy of this approximation.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

4.1: Introduction to Normal Distributions

Learning Objectives

- Describe the shape of normal distributions
- State 7 features of normal distributions

The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the "bell curve," although the tonal qualities of such a bell would be less than pleasing. It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss. As you will see in the section on the history of the normal distribution, although Gauss played an important role in its history, Abraham de Moivre first discovered the normal distribution.

Strictly speaking, it is not correct to talk about "the normal distribution" since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations. Figure $4.1.1$ shows three normal distributions. The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5, the distribution in red (the middle distribution) has a mean of 0 and a standard deviation of 1, and the distribution in black (right-most) has a mean of 2 and a standard deviation of 3. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

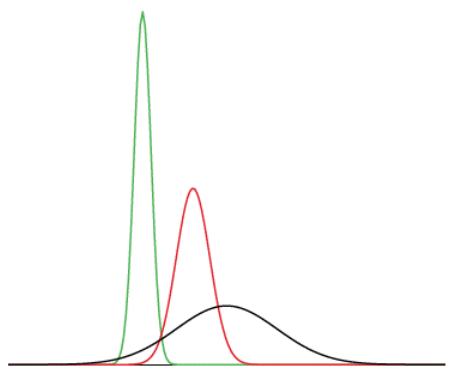


Figure $4.1.1$: Normal distributions differing in mean and standard deviation

The density of the normal distribution (the height for a given value on the x axis) is shown below. The parameters μ and σ are the mean and standard deviation, respectively, and define the normal distribution. The symbol e is the base of the natural logarithm and π is the constant pi.

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{4.1.1}$$

Since this is a non-mathematical treatment of statistics, do not worry if this expression confuses you. We will not be referring back to it in later sections.

Seven features of normal distributions are listed below. These features are illustrated in more detail in the remaining sections of this chapter.

1. Normal distributions are symmetric around their mean
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.

7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

4.2: Areas Under Normal Distributions

Learning Objectives

- State the proportion of a normal distribution within 1 and within 2 standard deviations of the mean
- Use the calculator "Calculate Area for a given X"
- Use the calculator "Calculate X for a given Area"

Areas under portions of a normal distribution can be computed by using calculus. Since this is a non-mathematical treatment of statistics, we will rely on computer programs and tables to determine these areas. Figure $4.2.1$ shows a normal distribution with a mean of 50 and a standard deviation of 10. The shaded area between 40 and 60 contains 68% of the distribution.

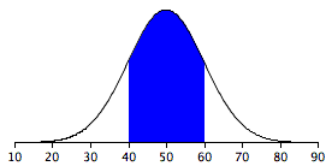


Figure $4.2.1$: *Normal distribution with a mean of* 50 *and standard deviation of* 10.

68% *of the area is within one standard deviation (*10*) of the mean (*50*)*

Figure $4.2.2$ shows a normal distribution with a mean of 100 and a standard deviation of 20. As in Figure $4.2.1$, 68% of the distribution is within one standard deviation of the mean.

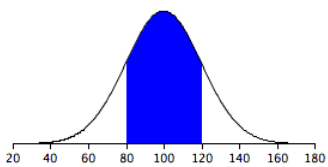


Figure $4.2.2$: *Normal distribution with a mean of* 100 *and standard deviation of* 20.

68% *of the area is within one standard deviation (*20*) of the mean (*100*).*

The normal distributions shown in Figures $4.2.1$ and $4.2.2$ are specific examples of the general rule that 68% of the area of any normal distribution is within one standard deviation of the mean.

Figure $4.2.3$ shows a normal distribution with a mean of 75 and a standard deviation of 10. The shaded area contains 95% of the area and extends from 55.4 to 94.6. For all normal distributions, 95% of the area is within 1.96 standard deviations of the mean. For quick approximations, it is sometimes useful to round off and use 2 rather than 1.96 as the number of standard deviations you need to extend from the mean so as to include 95% of the area.

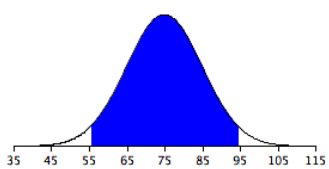


Figure $4.2.3$: *A normal distribution with a mean of* 75 *and a standard deviation of* 10.95% *of the area is within* 1.96 *standard deviations of the mean.*

The normal calculator can be used to calculate areas under the normal distribution. For example, you can use it to find the proportion of a normal distribution with a mean of 90 and a standard deviation of 12 that is above 110. Set the mean to 90 and the standard deviation to 12. Then enter "110" in the box to the right of the radio button "Above." At the bottom of the display you will see that the shaded area is 0.0478. See if you can use the calculator to find that the area between 115 and 120 is 0.0124

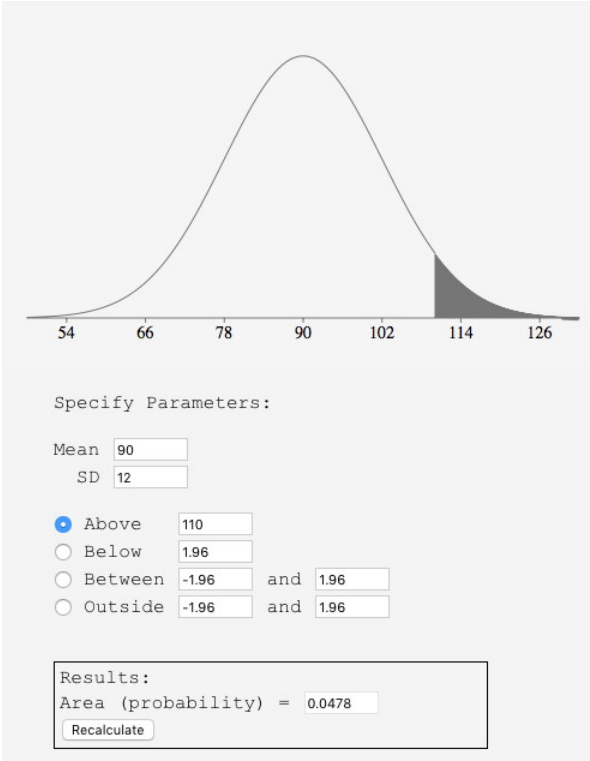


Figure 4.2.4: *Display from calculator showing the area above* 110.

Say you wanted to find the score corresponding to the 75th percentile of a normal distribution with a mean of 90 and a standard deviation of 12. Using the inverse normal calculator, you enter the parameters as shown in Figure 4.2.5 and find that the area below 98.09 is 0.75.

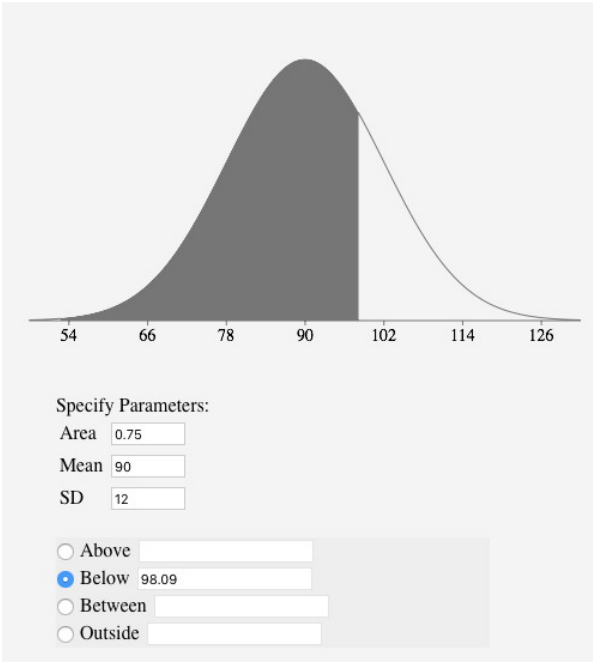


Figure 4.2.5: *Display from normal calculator showing that the* 75th *percentile is* 98.09.

Normal and Inverse Normal Calculator

- **7.3: Areas Under Normal Distributions** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

4.3: Varieties Demonstration

Learning Objectives

- Be able to describe differences in distributions based on their shape and statistics

Instructions

This demonstration allows you to change the mean and standard deviation of two normal distributions and observe the effects on the shapes of the distributions.When the demonstration begins, the red distribution has a mean of 50 and a standard deviaton of 10. The blue distribution has a mean of 50 and a standard deviation of 5. Notice how much more spread out the red distribution is than the blue distribution.

Choose different values for the two distributions and investigate their effects.

Illustrated Instructions

The demonstration starts by displaying two curves both with a mean of 50 and with standard deviations of 10 and 5 respectively. You can change the values for the mean and standard deviation for both curves.

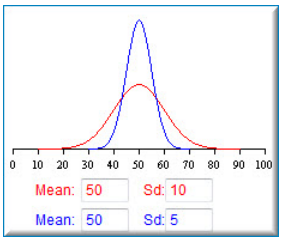


Figure $4.3.1$: Beginning curves for varieties demonstration

The screenshot below shows the distributions with different means and standard deviations. Note that the curves are on an axis that ranges from $0-100$, keep this in mind when changing the values of the mean and standard deviations as some values will not be displayed correctly.

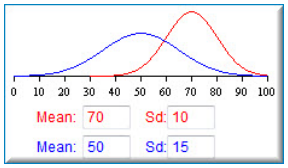


Figure $4.3.2$: Distributions with different means and standard deviations

4.4: Standard Normal Distribution

Learning Objectives

- State the mean and standard deviation of the standard normal distribution
- Use a Z table
- Use the normal calculator
- Transform raw data to Z scores

As discussed in the introductory section, normal distributions do not necessarily have the same means and standard deviations. A normal distribution with a mean of 0 and a standard deviation of 1 is called a standard normal distribution.

Areas of the normal distribution are often represented by tables of the standard normal distribution. A portion of a table of the standard normal distribution is shown in Table $4.4.1$.

Table $4.4.1$: A portion of a table of the standard normal distribution

Z	Area below
-2.5	0.0062
-2.49	0.0064
-2.48	0.0066
-2.47	0.0068
-2.46	0.0069
-2.45	0.0071
-2.44	0.0073
-2.43	0.0075
-2.42	0.0078
-2.41	0.008
-2.4	0.0082
-2.39	0.0084
-2.38	0.0087
-2.37	0.0089
-2.36	0.0091
-2.35	0.0094
-2.34	0.0096
-2.33	0.0099
-2.32	0.0102

The first column titled "Z" contains values of the standard normal distribution; the second column contains the area below Z. Since the distribution has a mean of 0 and a standard deviation of 1, the Z column is equal to the number of standard deviations below (or above) the mean. For example, a Z of -2.5 represents a value 2.5 standard deviations below the mean. The area below Z is 0.0062

The same information can be obtained using the following Java applet. Figure $4.4.1$ shows how it can be used to compute the area below a value of -2.5 on the standard normal distribution. Note that the mean is set to 0 and the standard deviation is set to 1.

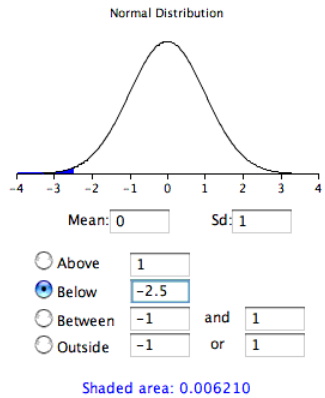


Figure 4.4.1: An example from the applet

Calculate Areas

A value from any normal distribution can be transformed into its corresponding value on a standard normal distribution using the following formula:

$$Z = \frac{X - \mu}{\sigma} \tag{4.4.1}$$

where Z is the value on the standard normal distribution, X is the value on the original distribution, μ is the mean of the original distribution, and σ is the standard deviation of the original distribution.

Example 4.4.1

As a simple application, what portion of a normal distribution with a mean of 50 and a standard deviation of 10 is below 26?

Solution

Applying the formula, we obtain

$$Z = \frac{26 - 50}{10} = -2.4 \tag{4.4.2}$$

From Table $4.4.1$, we can see that 0.0082 of the distribution is below -2.4. There is no need to transform to Z if you use the applet as shown in Figure $4.4.2$.

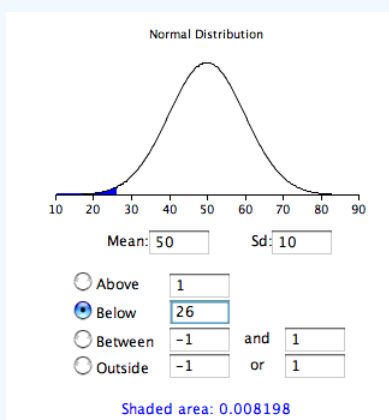


Figure 4.4.2: Area below 26 in a normal distribution with a mean of 50 and a standard deviation of 10

If all the values in a distribution are transformed to Z scores, then the distribution will have a mean of 0 and a standard deviation of 1. This process of transforming a distribution to one with a mean of 0 and a standard deviation of 1 is called standardizing the distribution.

4.5: Normal Approximation Demonstration

Learning Objectives

- State the relationship between sample size and the accuracy of normal approximation of the binomial distribution

Instructions

The normal distribution can be used to approximate the binomial distribution. This demonstration allows you to explore the accuracy of the approximation under a variety of conditions.

Illustrated Instructions

The demonstration displays the probability of success over a specific number of trials based on the entered total number of trials (N) and the probability of success on a given trial. You can change N as well as p and then select the proportion for which you would like the probability to be calculated.

By clicking the appropriate radio button, you can choose to calculate the probability above a specified value or between two values. Try out various values of N and p to calculate various approximations to the binomial to explore the accuracy of the approximation.

The calculation based on the normal approximation to the binomial is shown in green below and is equal to 0.1714 The actual binomial probability of $0.1719$is shown in red.

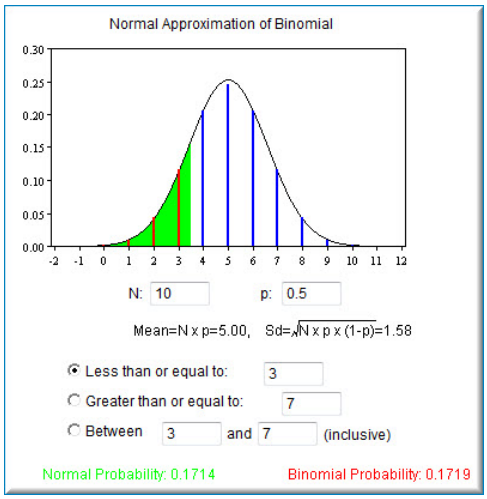


Figure $4.5.1$: Normal approximation demonstration

The screenshot below displays results for the probability of greater than 10 successful trials with 15 total trials and a 0.5 probability of success. The probability based on the normal approximation is displayed at the bottom of the screen in green next to the binomial probability displayed in red.

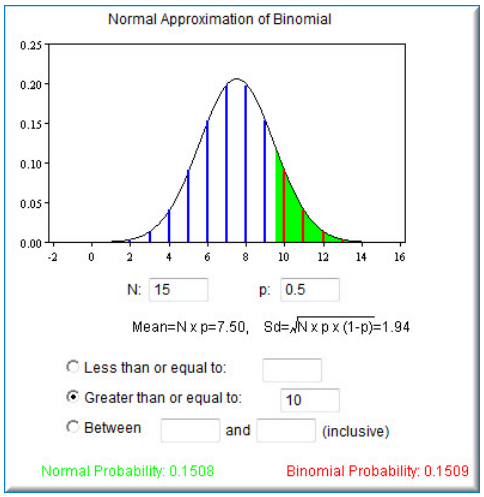


Figure 4.5.2: Normal approximation demonstration

4.6: Statistical Literacy

Learning Objectives

- Evaluating "Tail Risk"

Risk analyses often are based on the assumption of normal distributions. Critics have said that extreme events in reality are more frequent than would be expected assuming normality. The assumption has even been called a "Great Intellectual Fraud."

A recent article discussing how to protect investments against extreme events defined "tail risk" as "A tail risk, or extreme shock to financial markets, is technically defined as an investment that moves more than three standard deviations from the mean of a normal distribution of investment returns."

Example 4.6.1: what do you think?

Tail risk can be evaluated by assuming a normal distribution and computing the probability of such an event. Is that how "tail risk" should be evaluated?

Solution

Events more than three standard deviations from the mean are very rare for normal distributions. However, they are not as rare for other distributions such as highly-skewed distributions. If the normal distribution is used to assess the probability of tail events defined this way, then the "tail risk" will be underestimated.

4.E: Normal Distribution (Exercises)

You may want to use the "Calculate Area for a given X" and the "Calculate X for a given Area" calculators for some of these exercises.

General Questions

Q1

If scores are normally distributed with a mean of 35 and a standard deviation of 10, what percent of the scores is: (relevant section)

a. greater than 34
b. smaller than 42
c. between 28 and 34

Q2

a. What are the mean and standard deviation of the standard normal distribution?
b. What would be the mean and standard deviation of a distribution created by multiplying the standard normal distribution by 8 and then adding 75? (relevant section & here)

Q3

The normal distribution is defined by two parameters. What are they? (relevant section)

Q4

a. What proportion of a normal distribution is within one standard deviation of the mean?
b. What proportion is more than 2.0 standard deviations from the mean?
c. What proportion is between 1.25 and 2.1 standard deviations above the mean? (relevant section)

Q5

A test is normally distributed with a mean of 70 and a standard deviation of 8.

a. What score would be needed to be in the 85th percentile?
b. What score would be needed to be in the 22nd percentile? (relevant section)

Q6

Assume a normal distribution with a mean of 70 and a standard deviation of 12. What limits would include the middle 65% of the cases? (relevant section)

Q7

A normal distribution has a mean of 20 and a standard deviation of 4. Find the Z scores for the following numbers: (relevant section)

a. 28
b. 18
c. 10
d. 23

Q8

Assume the speed of vehicles along a stretch of $I-10$ has an approximately normal distribution with a mean of 71 mph and a standard deviation of 8 mph.

a. The current speed limit is 65 mph. What is the proportion of vehicles less than or equal to the speed limit?
b. What proportion of the vehicles would be going less than 50 mph?
c. A new speed limit will be initiated such that approximately 10% of vehicles will be over the speed limit. What is the new speed limit based on this criterion?
d. In what way do you think the actual distribution of speeds differs from a normal distribution? (relevant section)

Q9

A variable is normally distributed with a mean of 120 and a standard deviation of 5. One score is randomly sampled. What is the probability it is above 127? (relevant section)

Q10

You want to use the normal distribution to approximate the binomial distribution. Explain what you need to do to find the probability of obtaining exactly 7 heads out of 12 flips. (relevant section)

Q11

A group of students at a school takes a history test. The distribution is normal with a mean of 25, and a standard deviation of 4.

a. Everyone who scores in the top 30% of the distribution gets a certificate. What is the lowest score someone can get and still earn a certificate?
b. The top 5% of the scores get to compete in a statewide history contest. What is the lowest score someone can get and still go onto compete with the rest of the state? (relevant section)

Q12

Use the normal distribution to approximate the binomial distribution and find the probability of getting 15 to 18 heads out of 25 flips. Compare this to what you get when you calculate the probability using the binomial distribution. Write your answers out to four decimal places. (relevant section & relevant section)

Q13

True/false: For any normal distribution, the mean, median, and mode will have the same value. (relevant section)

Q14

True/false: In a normal distribution, 11.5% of scores are greater than $Z = 1.2$. (relevant section)

Q15

True/false: The percentile rank for the mean is 50% for any normal distribution. (relevant section)

Q16

True/false: The larger the π, the better the normal distribution approximates the binomial distribution. (relevant section & relevant section)

Q17

True/false: A Z-score represents the number of standard deviations above or below the mean. (relevant section)

Q18

True/false: Abraham de Moivre, a consultant to gamblers, discovered the normal distribution when trying to approximate the binomial distribution to make his computations easier. (relevant section)

Q19

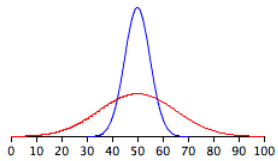
True/false: The standard deviation of the blue distribution shown below is about 10. (relevant section)

Q20

True/false: In the figure below, the red distribution has a larger standard deviation than the blue distribution. (relevant section)

Q21

True/false: The red distribution has more area underneath the curve than the blue distribution does. (relevant section)



Questions from Case Studies

The following question uses data from the Angry Moods (AM) case study.

Q22

For this problem, use the Anger Expression (AE) scores.

a. Compute the mean and standard deviation.
b. Then, compute what the 25th, 50th and 75th percentiles would be if the distribution were normal.
c. Compare the estimates to the actual 25th, 50th and 75th percentiles. (relevant section)

The following question uses data from the Physicians Reaction (PR) case study.

Q23

For this problem, use the time spent with the overweight patients.

a. Compute the mean and standard deviation of this distribution.
b. What is the probability that if you chose an overweight participant at random, the doctor would have spent 31 minutes or longer with this person?
c. Now assume this distribution is normal (and has the same mean and standard deviation). Now what is the probability that if you chose an overweight participant at random, the doctor would have spent 31 minutes or longer with this person? (relevant section)

The following questions are from ARTIST (reproduced with permission)

Assessment Resource Tools for Improving Statistical Thinking

Visit the site

Q24

A set of test scores are normally distributed. Their mean is 100 and standard deviation is 20. These scores are converted to standard normal z scores. What would be the mean and median of this distribution?

a. 0
b. 1
c. 50
d. 100

Q25

Suppose that weights of bags of potato chips coming from a factory follow a normal distribution with mean 12.8 ounces and standard deviation 0.6 ounces. If the manufacturer wants to keep the mean at 12.8 ounces but adjust the standard deviation so that only 1% of the bags weigh less than 12 ounces, how small does he/she need to make that standard deviation?

Q26

A student received a standardized (z) score on a test that was -0.57. What does this score tell about how this student scored in relation to the rest of the class? Sketch a graph of the normal curve and shade in the appropriate area.

Q27

Suppose you take 50 measurements on the speed of cars on Interstate 5, and that these measurements follow roughly a Normal distribution. Do you expect the standard deviation of these 50 measurements to be about 1 mph, 5 mph, 10 mph, or 20 mph? Explain.

Q28

Suppose that combined verbal and math SAT scores follow a normal distribution with mean 896 and standard deviation 174. Suppose further that Peter finds out that he scored in the top 3% of SAT scores. Determine how high Peter's score must have been.

Q29

Heights of adult women in the United States are normally distributed with a population mean of $\mu = 63.5$ inches and a population standard deviation of $\sigma = 2.5$. A medical researcher is planning to select a large random sample of adult women to participate in a future study. What is the standard value, or z-value, for an adult woman who has a height of 68.5 inches?

Q30

An automobile manufacturer introduces a new model that averages 27 miles per gallon in the city. A person who plans to purchase one of these new cars wrote the manufacturer for the details of the tests, and found out that the standard deviation is 3 miles per gallon. Assume that in-city mileage is approximately normally distributed.

a. What is the probability that the person will purchase a car that averages less than 20 miles per gallon for in-city driving?
b. What is the probability that the person will purchase a car that averages between 25 and 29 miles per gallon for in-city driving?

Select Answers

S1

b. 75.8%

S2

b. Mean = 75

S4

c. 0.088

S5

a. 78.3

S7

a. 2.0

S8

a. 0.227

S11

a. 27.1

S12

0.2037(normal approximation)

S22

25th percentile:

b. 28.27
c. 26.75

S23

b. 0.053

CHAPTER OVERVIEW

5: Sampling Distributions

The concept of a sampling distribution is perhaps the most basic concept in inferential statistics. It is also a difficult concept because a sampling distribution is a theoretical distribution rather than an empirical distribution. The introductory section defines the concept and gives an example for both a discrete and a continuous distribution. It also discusses how sampling distributions are used in inferential statistics. The Basic Demo is an interactive demonstration of sampling distributions. It is designed to make the abstract concept of sampling distributions more concrete. The Sample Size Demo allows you to investigate the effect of sample size on the sampling distribution of the mean. The Central Limit Theorem (CLT) Demo is an interactive illustration of a very important and counter-intuitive characteristic of the sampling distribution of the mean. The remaining sections of the chapter concern the sampling distributions of important statistics: the Sampling Distribution of the Mean, the Sampling Distribution of the Difference Between Means, the Sampling Distribution of r, and the Sampling Distribution of a Proportion.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

5.1: Introduction to Sampling Distributions

Learning Objectives

- Define inferential statistics
- Graph a probability distribution for the mean of a discrete variable
- Describe a sampling distribution in terms of "all possible outcomes"
- Describe a sampling distribution in terms of repeated sampling
- Describe the role of sampling distributions in inferential statistics
- Define the standard error of the mean

Suppose you randomly sampled 10 people from the population of women in Houston, Texas, between the ages of 21 and 35 years and computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in Houston. It might be somewhat lower or it might be somewhat higher, but it would not equal the population mean exactly. Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.

Recall that inferential statistics concern generalizing from a sample to a population. A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter. (In this example, the sample statistics are the sample means and the population parameter is the population mean.) As the later portions of this chapter show, these determinations are based on sampling distributions.

Discrete Distributions

We will illustrate the concept of sampling distributions with a simple example. Figure 5.1.1 shows three pool balls, each with a number on it. Two of the balls are selected randomly (with replacement) and the average of their numbers is computed.

Figure 5.1.1: *The pool balls*

All possible outcomes are shown below in Table 5.1.1.

Table 5.1.1: *All possible outcomes when two balls are sampled with replacement*

Outcome	Ball 1	Ball 2	Mean
1	1	1	1.0
2	1	2	1.5
3	1	3	2.0
4	2	1	1.5
5	2	2	2.0
6	2	3	2.5
7	3	1	2.0
8	3	2	2.5
9	3	3	3.0

Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0. The frequencies of these means are shown in Table 5.1.2. The relative frequencies are equal to the frequencies divided by nine because there are nine possible outcomes.

Table 5.1.2: *Frequencies of means for* $N = 2$

Mean	Frequency	Relative Frequency
1.0	1	0.111
1.5	2	0.222
2.0	3	0.333
2.5	2	0.222
3.0	1	0.111

Figure 5.1.2 shows a relative frequency distribution of the means based on Table 5.1.2. This distribution is also a probability distribution since the Y-axis is the probability of obtaining a given mean from a sample of two balls in addition to being the relative frequency.

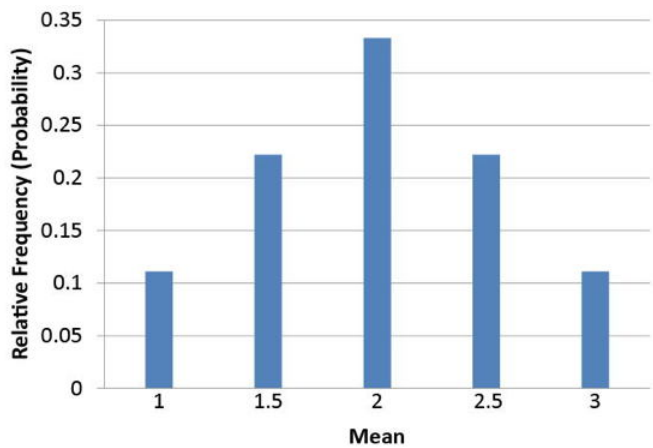


Figure 5.1.2: *Distribution of means for* $N = 2$

The distribution shown in Figure 5.1.2 is called the sampling distribution of the mean. Specifically, it is the sampling distribution of the mean for a sample size of 2 ($N = 2$). For this simple example, the distribution of pool balls and the sampling distribution are both discrete distributions. The pool balls have only the values 1, 2, and 3, and a sample mean can have one of only five values shown in Table 5.1.2.

There is an alternative way of conceptualizing a sampling distribution that will be useful for more complex distributions. Imagine that two balls are sampled (with replacement) and the mean of the two balls is computed and recorded. Then this process is repeated for a second sample, a third sample, and eventually thousands of samples. After thousands of samples are taken and the mean computed for each, a relative frequency distribution is drawn. The more samples, the closer the relative frequency distribution will come to the sampling distribution shown in Figure 5.1.2. As the number of samples approaches infinity, the relative frequency distribution will approach the sampling distribution. This means that you can conceive of a sampling distribution as being a relative frequency distribution based on a very large number of samples. To be strictly correct, the relative frequency distribution approaches the sampling distribution as the number of samples approaches infinity.

Table 5.1.3: *All possible outcomes when two balls are sampled with replacement*

Outcome	Ball 1	Ball 2	Range
1	1	1	0
2	1	2	1
3	1	3	2
4	2	1	1
5	2	2	0

Outcome	Ball 1	Ball 2	Range
6	2	3	1
7	3	1	2
8	3	2	1
9	3	3	0

It is important to keep in mind that every statistic, not just the mean, has a sampling distribution. For example, Table $5.1.3$ shows all possible outcomes for the range of two numbers (larger number minus the smaller number). Table $5.1.4$ shows the frequencies for each of the possible ranges and Figure $5.1.3$ shows the sampling distribution of the range.

Table $5.1.4$: *Frequencies of ranges for* $N = 2$

Range	Frequency	Relative Frequency
0	3	0.333
1	4	0.444
2	2	0.222

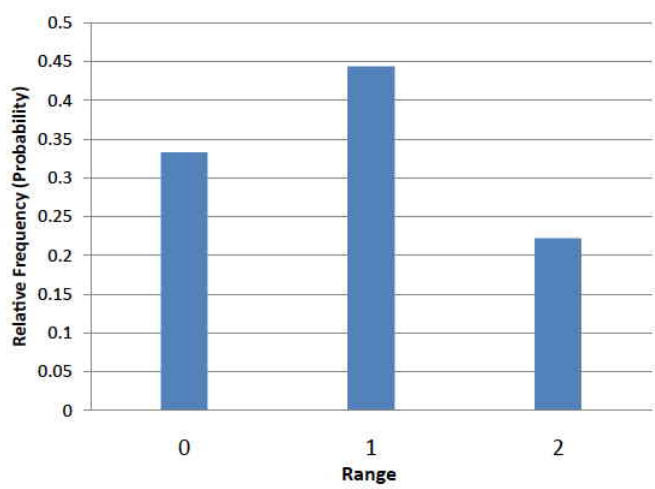


Figure $5.1.3$: *Distribution of ranges for* $N = 2$

It is also important to keep in mind that there is a sampling distribution for various sample sizes. For simplicity, we have been using $N = 2$. The sampling distribution of the range for $N = 3$ is shown in Figure $5.1.4$.

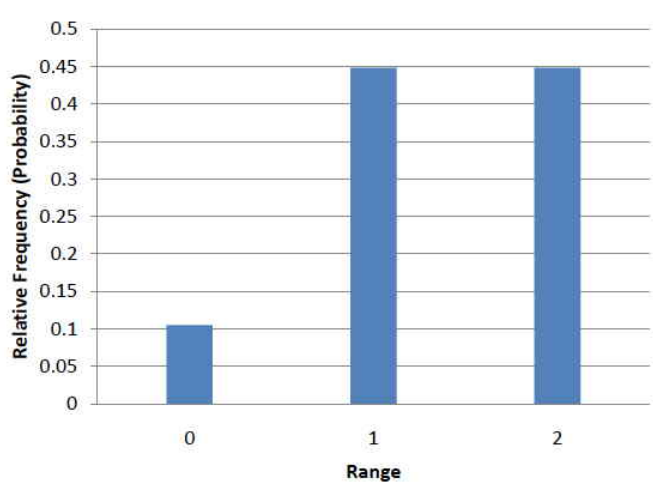


Figure $5.1.4$: *Distribution of ranges for* $N = 3$

Continuous Distributions

In the previous section, the population consisted of three pool balls. Now we will consider sampling distributions when the population distribution is continuous. What if we had a thousand pool balls with numbers ranging from 0.001 to 1.000 in equal steps? (Although this distribution is not really continuous, it is close enough to be considered continuous for practical purposes.) As before, we are interested in the distribution of means we would get if we sampled two balls and computed the mean of these two balls. In the previous example, we started by computing the mean for each of the nine possible outcomes. This would get a bit tedious for this example since there are $1,000,000$possible outcomes ($1,000$ for the first ball x $1,000$ for the second). Therefore, it is more convenient to use our second conceptualization of sampling distributions which conceives of sampling distributions in terms of relative frequency distributions. Specifically, the relative frequency distribution that would occur if samples of two balls were repeatedly taken and the mean of each sample computed.

When we have a truly continuous distribution, it is not only impractical but actually impossible to enumerate all possible outcomes. Moreover, in continuous distributions, the probability of obtaining any single value is zero. Therefore, as discussed in the section "Introduction to Distributions," these values are called probability densities rather than probabilities.

Sampling Distributions and Inferential Statistics

As we stated in the beginning of this chapter, sampling distributions are important for inferential statistics. In the examples given so far, a population was specified and the sampling distribution of the mean and the range were determined. In practice, the process proceeds the other way: you collect sample data and from these data you estimate parameters of the sampling distribution. This knowledge of the sampling distribution can be very useful. For example, knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean. Fortunately, this information is directly available from a sampling distribution. The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the standard error of the mean. If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

To be specific, assume your sample mean were 125 and you estimated that the standard error of the mean were 5 (using a method shown in a later section). If you had a normal distribution, then it would be likely that your sample mean would be within 10 units of the population mean since most of a normal distribution is within two standard deviations of the mean.

Keep in mind that all statistics have sampling distributions, not just the mean. In later sections we will be discussing the sampling distribution of the variance, the sampling distribution of the difference between means, and the sampling distribution of Pearson's correlation, among others.

5.2: Basic Demo

Learning Objectives

- Develop a basic understanding of the properties of a sampling distribution based on the properties of the population

Instructions

This simulation illustrates the concept of a sampling distribution. Depicted on the top graph is the population from which we are going to sample. There are 33 different values in the population: the integers from 0 to 32 (inclusive). You can think of the population as consisting of having an extremely large number of balls with $0's$, an extremely large number with $1's$, etc. on them. The height of the distribution shows the relative number of balls of each number. There is an equal number of balls for each number, so the distribution is a rectangle.

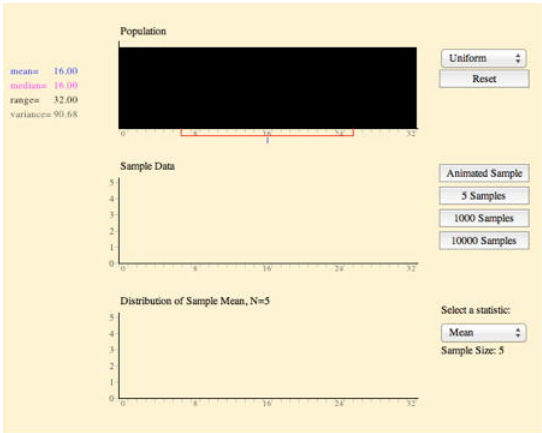


Figure $5.2.1$: Sample Distribution Demonstration

If you push the "animated sampling" button, five balls are selected and and are plotted on the second graph. The mean of this sample of five is then computed and plotted on the third graph. If you push the "animated sampling" button again, another sample of five will be taken, and again plotted on the second graph. The mean will be computed and plotted on the third graph. This third graph is labeled "Distribution of Sample Means, $N = 5$" because each value plotted is a sample mean based on a sample of five. At this point, you should have two means plotted in this graph.

The mean is depicted graphically on the distributions themselves by a blue vertical bar below the X-axis. A red line starts from this mean value and extends one standard deviation in length in both directions. The values of both the mean and the standard deviation are given to the left of the graph. Notice that the numeric form of a property matches its graphical form.

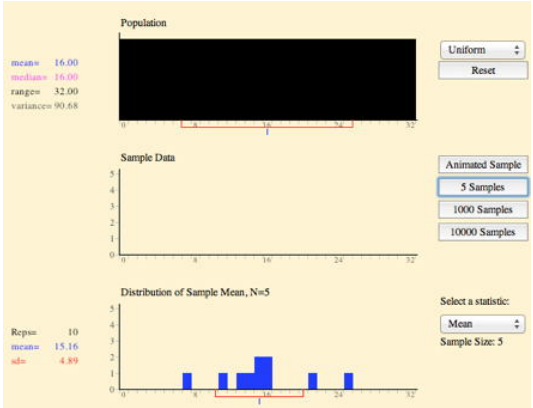


Figure $5.2.2$: Sample Distribution Demonstration

The sampling distribution of a statistic is the relative frequency distribution of that statistic that is approached as the number of samples (not the sample size!) approaches infinity. To approximate a sampling distribution, click the "$5,000$ samples" button several times. The bottom graph is then a relative frequency distribution of the thousands of means. It is not truly a sampling distribution because it is based on a finite number of samples. Nonetheless, it is a very good approximation.

The simulation has been explained in terms of the sampling distribution of the mean for $N = 5$. All statistics, not just the mean, have sampling distributions. Moreover, there is a different sampling distribution for each value of N. For the sake of simplicity, this simulation only uses $N = 5$. Finally, the default is to sample from a distribution for which each value has an equal chance of occurring. Other shapes of the distribution are possible. In this simulation, you can make the population normally distributed as well.

In this simulation, you can specify a sample statistic (the default is mean) and then sample a sufficiently large number of samples until the sampling distribution stabilizes. Make sure you understand the difference between the sample size (which here is 5) and the number of samples included in a distribution.

You should also compare the value of a statistic in the population and the mean of the sampling distribution of that statistic. For some statistics, the mean of the sampling distribution will be very close to the corresponding population parameter; for at least one, there will be a large difference. Also note how the overall shape of sampling distribution differs from that of the population.

Finally, as shown in the video demo, you can change the parent population to a normal distribution from a uniform distribution.

Video Instructions

Video Demo

The video above demonstrates the use of the Sampling distribution Demonstration. The first graph represents the distribution of the population from which the sample will be drawn. In the video this distribution is changed to normal. Each time the "Animated Sample" button is clicked a random sample of five elements is drawn from the population. You can draw multiple samples of 5 by clicking on the buttons directly below "Animated Sample". The mean of each of these sample is displayed in the third graph on at the bottom. The graph can also be set to display other descriptive statistics besides the mean.

If you look to the left of this third graph you can see the mean and standard deviation of the sampling distribution. Try drawing $50,000$ samples from both types of population distributions and compare the sampling distribution statistics to their equivalent population statistics to see if you can discover any trends.

5.3: Sample Size Demo

Learning Objectives

- Compare the sampling distributions of different sample sizes

Instructions

This simulation demonstrates the effect of sample size on the sampling distribution.

Depicted on the top graph is the population distribution. By default it is a uniform distribution (all values are equally likely). The sampling distributions for two different sample sizes are shown in the lower two graphs. The starting values are 2 and 10. By default, the statistic to be computed is the mean, although you can also specify to compute the median. For both the population distribution and the sampling distribution, the mean and the standard deviation are depicted graphically on the frequency distribution itself. The blue-colored vertical bar below the X-axis indicates the mean value. The red line starts from this mean value and extends one standard deviation in length in both directions. The values of both the mean and the standard deviation are also given to the left of the graph. Notice that the numeric form of a property matches its graphical form in color. In this simulation, you specify two sample sizes (the defaults are set at $N = 2$ and $N = 10$), and then sample a sufficiently large number of samples until the sampling distributions stabilize. Compare the mean and standard deviaiton of the two sampling distributions. Repeat the process a couple times and watch the results. Do you observe a general rule regarding the effect of sample size on the mean and the standard deviation of the sampling distribution? You may also test the effect of sample size with a normal population or with a different sample statistic (the median). When you have discovered the rule, go back and answer the questions again.

Illustrated Instructions

Video Demonstration

The video below changes the population distribution from uniform to normal and then draws $30,000$ samples with $N = 2$ and $30,000$ samples with $N = 10$ by clicking the "$10,000$ Samples" 3 times. Notice the differences in the means and standard deviations of the two sample distributions. How do these compare to the population?

The vertical bar to the right of each sampling distribution can be dragged along the x-axis and once the mouse is released the area of the curve to the left of the line is displayed above the chart.

For some browsers you will not see the bars move as you move them. They will move when you release the mouse button.

5.4: Central Limit Theorem Demonstration

Learning Objectives

- Develop a basic understanding of the properties of a sampling distribution based on the properties of the population

Instructions

This simulation demonstrates the effect of sample size on the shape of the sampling distribution of the mean.

Depicted on the top graph is the population which is sometimes referred to as the parent distribution. Two sampling distributions of the mean, associated with their respective sample size will be created on the second and third graphs.

For both the population distribution and the sampling distributions, their mean and the standard deviation are depicted graphically on the frequency distribution itself. The blue-colored vertical bar below the X-axis indicates where the mean value falls. The red line starts from this mean value and extends one standard deviation in length in both directions. The values of both the mean and the standard deviation are also given to the left of the graph. Notice that the numeric form of a property matches its graphical form in color. In addition, the skew and the kurtosis of each distribution are also provided to the left. These two variables are determined by the shape of distribution. The skew and kurtosis for a normal distribution are both 0.

In this simulation, you need to first specify a population (the default is uniform distribution). Take note of the skew and kurtosis of the population. Then pick two different sample sizes (the defaults are $N = 2$ and $N = 10$), and sample a sufficiently large number of samples until the sampling distributions change relatively little with additional samples (about $50,000$ samples.) Observe the overall shape of the two sampling distributions, and further compare their means, standard deviations, skew and kurtosis. Change the sample sizes and repeat the process a few times. Do you observe a general rule regarding the effect of sample size on the shape of the sampling distribution?

You may also test the effect of sample size with populations of other shape (uniform, skewed or custom ones).

Illustrated Instructions

Central Limit Theorem Video Demo

The video below changes the population distribution to skewed and draws $100,000$ samples with $N = 2$ and $N = 10$ with the " $10,000$ Samples" button. Note the statistics and shape of the two sample distributions how do these compare to each other and to the population?

5.5: Sampling Distribution of the Mean

Learning Objectives

- State the mean and variance of the sampling distribution of the mean
- Compute the standard error of the mean
- State the central limit theorem

The sampling distribution of the mean was defined in the section introducing sampling distributions. This section reviews some important properties of the sampling distribution of the mean introduced in the demonstrations in this chapter.

Sampling Mean

The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean is also μ. The symbol μ_M is used to refer to the mean of the sampling distribution of the mean. Therefore, the formula for the mean of the sampling distribution of the mean can be written as:

$$\mu_M = \mu \tag{5.5.1}$$

Sampling Variance

The variance of the sampling distribution of the mean is computed as follows:

$$\sigma_M^2 = \frac{\sigma^2}{N} \tag{5.5.2}$$

That is, the variance of the sampling distribution of the mean is the population variance divided by N, the sample size (the number of scores used to compute a mean). Thus, the larger the sample size, the smaller the variance of the sampling distribution of the mean.

Note

(optional) This expression can be derived very easily from the variance sum law. Let's begin by computing the variance of the sampling distribution of the sum of three numbers sampled from a population with variance σ^2. The variance of the sum would be:

$$\sigma^2 + \sigma^2 + \sigma^2 \tag{5.5.3}$$

For N numbers, the variance would be $N\sigma^2$. Since the mean is $1/N$ times the sum, the variance of the sampling distribution of the mean would be $1/N^2$ times the variance of the sum, which equals σ^2/N.

The standard error of the mean is the standard deviation of the sampling distribution of the mean. It is therefore the square root of the variance of the sampling distribution of the mean and can be written as:

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{5.5.4}$$

The standard error is represented by a σ because it is a standard deviation. The subscript (M) indicates that the standard error in question is the standard error of the mean.

Central Limit Theorem

Definition

The central limit theorem states that:

Given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2/N as N, the sample size, increases.

The expressions for the mean and variance of the sampling distribution of the mean are not new or remarkable. What is remarkable is that regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as N increases. If you have used the "Central Limit Theorem Demo," you have already seen this for yourself. As a reminder, Figure $5.5.1$ shows the results of the simulation for $N = 2$ and $N = 10$. The parent population was a uniform distribution. You can see that the distribution for $N = 2$ is far from a normal distribution. Nonetheless, it does show that the scores are denser in the middle than in the tails. For $N = 10$ the distribution is quite close to a normal distribution. Notice that the means of the two distributions are the same, but that the spread of the distribution for $N = 10$ is smaller.

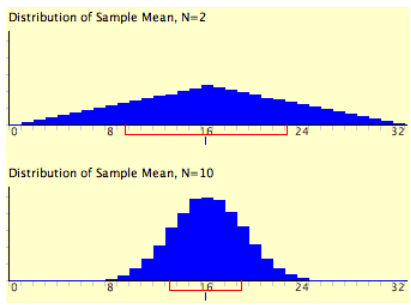


Figure $5.5.1$: *A simulation of a sampling distribution. The parent population is uniform. The blue line under "16" indicates that* 16 *is the mean. The red line extends from the mean plus and minus one standard deviation.*

Figure $5.5.2$ shows how closely the sampling distribution of the mean approximates a normal distribution even when the parent population is very non-normal. If you look closely you can see that the sampling distributions do have a slight positive skew. The larger the sample size, the closer the sampling distribution of the mean would be to a normal distribution.

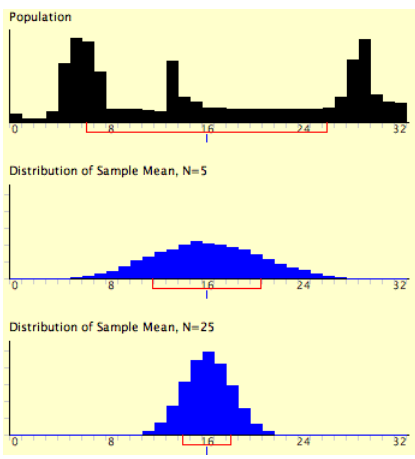


Figure $5.5.2$: *A simulation of a sampling distribution. The parent population is very non-normal.*

5.6: Difference Between Means

Learning Objectives

- State the mean and variance of the sampling distribution of the difference between means
- Compute the standard error of the difference between means
- Compute the probability of a difference between means being above a specified value

Statistical analyses are very often concerned with the difference between means. A typical example is an experiment designed to compare the mean of a control group with the mean of an experimental group. Inferential statistics used in the analysis of this type of experiment depend on the sampling distribution of the difference between means.

The sampling distribution of the difference between means can be thought of as the distribution that would result if we repeated the following three steps over and over again:

1. sample n_1 scores from Population 1 and n_2 scores from Population 2
2. compute the means of the two samples (M_1 and M_2)
3. compute the difference between means, $M_1 - M_2$

The distribution of the differences between means is the sampling distribution of the difference between means.

As you might expect, the mean of the sampling distribution of the difference between means is:

$$\mu_{M_1-M_2} = \mu_1 - \mu_2 \tag{5.6.1}$$

which says that the mean of the distribution of differences between sample means is equal to the difference between population means. For example, say that the mean test score of all 12-year-olds in a population is 34 and the mean of 10-year-olds is 25. If numerous samples were taken from each age group and the mean difference computed each time, the mean of these numerous differences between sample means would be $34 - 25 = 9$.

From the variance sum law, we know that:

$$\sigma^2_{M_1-M_2} = \sigma^2_{M_1} + \sigma^2_{M_2} \tag{5.6.2}$$

which says that the variance of the sampling distribution of the difference between means is equal to the variance of the sampling distribution of the mean for Population 1 plus the variance of the sampling distribution of the mean for Population 2. Recall the formula for the variance of the sampling distribution of the mean:

$$\sigma^2_M = \frac{\sigma^2}{N} \tag{5.6.3}$$

Since we have two populations and two samples sizes, we need to distinguish between the two variances and sample sizes. We do this by using the subscripts 1 and 2. Using this convention, we can write the formula for the variance of the sampling distribution of the difference between means as:

$$\sigma^2_{M_1-M_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{5.6.4}$$

Since the standard error of a sampling distribution is the standard deviation of the sampling distribution, the standard error of the difference between means is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{5.6.5}$$

Just to review the notation, the symbol on the left contains a sigma (σ), which means it is a standard deviation. The subscripts $M_1 - M_2$ indicate that it is the standard deviation of the sampling distribution of $M_1 - M_2$.

Now let's look at an application of this formula.

Example 5.6.1

Assume there are two species of green beings on Mars. The mean height of Species 1 is 32 while the mean height of Species 2 is 22. The variances of the two species are 60 and 70, respectively and the heights of both species are normally distributed. You randomly sample 10 members of Species 1 and 14 members of Species 2. What is the probability that the mean of the 10 members of Species 1 will exceed the mean of the 14 members of Species 2 by 5 or more? Without doing any calculations, you probably know that the probability is pretty high since the difference in population means is 10. But what exactly is the probability?

Solution

First, let's determine the sampling distribution of the difference between means. Using the formulas above, the mean is

$$\mu_{M_1-M_2} = 32 - 22 = 10 \tag{5.6.6}$$

The standard error is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{60}{10} + \frac{70}{14}} = 3.317 \tag{5.6.7}$$

The sampling distribution is shown in Figure 5.6.1. Notice that it is normally distributed with a mean of 10 and a standard deviation of 3.317. The area above 5 is shaded blue.

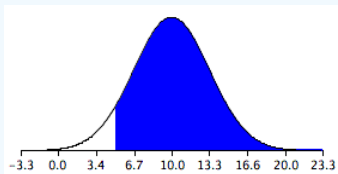


Figure 5.6.1: The sampling distribution of the difference between means

The last step is to determine the area that is shaded blue. Using either a Z table or the normal calculator, the area can be determined to be 0.934. Thus the probability that the mean of the sample from Species 1 will exceed the mean of the sample from Species 2 by 5 or more is 0.934.

As shown below, the formula for the standard error of the difference between means is much simpler if the sample sizes and the population variances are equal. When the variances and samples sizes are the same, there is no need to use the subscripts 1 and 2 to differentiate these terms.

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}} \tag{5.6.8}$$

This simplified version of the formula can be used for the following problem.

Example 5.6.2

The mean height of 15-year-old boys (in cm) is 175 and the variance is 64. For girls, the mean is 165 and the variance is 64. If eight boys and eight girls were sampled, what is the probability that the mean height of the sample of girls would be higher than the mean height of the sample of boys? In other words, what is the probability that the mean height of girls minus the mean height of boys is greater than 0?

Solution

As before, the problem can be solved in terms of the sampling distribution of the difference between means (girls - boys). The mean of the distribution is 165 - 175 = -10. The standard deviation of the distribution is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{2\sigma^2}{n}} = \sqrt{\frac{(2)(64)}{8}} = 4 \tag{5.6.9}$$

A graph of the distribution is shown in Figure 5.6.2. It is clear that it is unlikely that the mean height for girls would be higher than the mean height for boys since in the population boys are quite a bit taller. Nonetheless it is not inconceivable that the

girls' mean could be higher than the boys' mean.

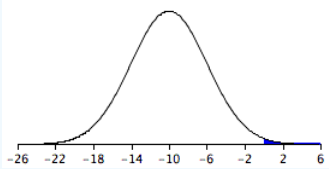


Figure 5.6.2: *sampling distribution of the difference between mean heights*

A difference between means of 0 or higher is a difference of $10/4 = 2.5$ standard deviations above the mean of -10. The probability of a score 2.5 or more standard deviations above the mean is 0.0062

5.7: Sampling Distribution of Pearson's r

Learning Objectives

- Transform r to z'
- Calculate the probability of obtaining an r above a specified value

Assume that the correlation between quantitative and verbal SAT scores in a given population is 0.60. In other words, $\rho = 0.60$. If 12 students were sampled randomly, the sample correlation, r, would not be exactly equal to 0.60. Naturally different samples of 12 students would yield different values of r. The distribution of values of r after repeated samples of 12 students is the sampling distribution of r.

The shape of the sampling distribution of r for the above example is shown in Figure $5.7.1$. You can see that the sampling distribution is not symmetric: it is negatively skewed. The reason for the skew is that r cannot take on values greater than 1.0 and therefore the distribution cannot extend as far in the positive direction as it can in the negative direction. The greater the value of ρ, the more pronounced the skew.

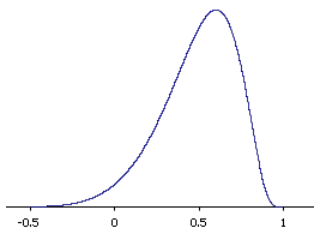


Figure $5.7.1$: The sampling distribution of r for $N = 12$ and $\rho = 0.60$

Figure $5.7.2$ shows the sampling distribution for $\rho = 0.90$. This distribution has a very short positive tail and a long negative tail.

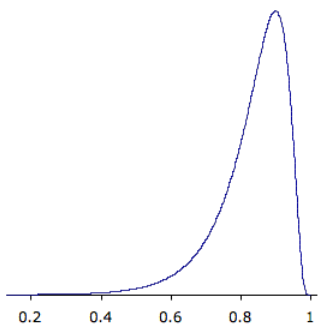


Figure $5.7.1$: The sampling distribution of r for $N = 12$ and $\rho = 0.90$

Referring back to the SAT example, suppose you wanted to know the probability that in a sample of 12 students, the sample value of r would be 0.75 or higher. You might think that all you would need to know to compute this probability is the mean and standard error of the sampling distribution of r. However, since the sampling distribution is not normal, you would still not be able to solve the problem. Fortunately, the statistician Fisher developed a way to transform r to a variable that is normally distributed with a known standard error. The variable is called z' and the formula for the transformation is given below.

$$z' = 0.5 \ \ln \frac{1+r}{1-r} \tag{5.7.1}$$

The details of the formula are not important here since normally you will use either a table, Pearson r to Fisher z' or calculator to do the transformation. What is important is that z' is normally distributed and has a standard error of

$$\frac{1}{\sqrt{N-3}} \tag{5.7.2}$$

where N is the number of pairs of scores.

Example 5.7.1

Let's return to the question of determining the probability of getting a sample correlation of 0.75 or above in a sample of 12 from a population with a correlation of 0.60.

Solution

The first step is to convert both 0.60 and 0.75 to their z' values, which are 0.693 and 0.973, respectively.

The standard error of z' for $N = 12$ is 0.333.

Therefore the question is reduced to the following: given a normal distribution with a mean of 0.693 and a standard deviation of 0.333, what is the probability of obtaining a value of 0.973 or higher?

The answer can be found directly from the applet "Calculate Area for a given X" to be 0.20.

Alternatively, you could use the formula:

$$z = \frac{X - \mu}{\sigma} = \frac{0.973 - 0.693}{0.333} = 0.841 \tag{5.7.3}$$

and use a table to find that the area above 0.841 is 0.20.

5.8: Sampling Distribution of p

Learning Objectives

- State the relationship between the sampling distribution of p and the normal distribution

Assume that in an election race between Candidate A and Candidate B, 0.60 of the voters prefer Candidate A. If a random sample of 10 voters were polled, it is unlikely that exactly 60% of them (6) would prefer Candidate A. By chance the proportion in the sample preferring Candidate A could easily be a little lower than 0.60 or a little higher than 0.60. The sampling distribution of p is the distribution that would result if you repeatedly sampled 10 voters and determined the proportion (p) that favored Candidate A.

The sampling distribution of p is a special case of the sampling distribution of the mean. Table 5.8.1 shows a hypothetical random sample of 10 voters. Those who prefer Candidate A are given scores of 1 and those who prefer Candidate B are given scores of 0. Note that seven of the voters prefer Candidate A so the sample proportion (p) is

$$p = \frac{7}{10} = 0.70 \tag{5.8.1}$$

As you can see, p is the mean of the 10 preference scores.

Table 5.8.1: Sample of voters

Voter	Preference
1	1
2	0
3	1
4	1
5	1
6	0
7	1
8	0
9	1
10	1

The distribution of p is closely related to the binomial distribution. The binomial distribution is the distribution of the total number of successes (favoring Candidate A, for example) whereas the distribution of p is the distribution of the mean number of successes. The mean, of course, is the total divided by the sample size, N. Therefore, the sampling distribution of p and the binomial distribution differ in that p is the mean of the scores (0.70) and the binomial distribution is dealing with the total number of successes (7).

The binomial distribution has a mean of

$$\mu = N\pi \tag{5.8.2}$$

Dividing by N to adjust for the fact that the sampling distribution of p is dealing with means instead of totals, we find that the mean of the sampling distribution of p is:

$$\mu_p = \pi \tag{5.8.3}$$

The standard deviation of the binomial distribution is:

$$\sqrt{N\pi(1-\pi)} \tag{5.8.4}$$

Dividing by N because p is a mean not a total, we find the standard error of p:

$$\sigma_p = \frac{\sqrt{N\pi(1-\pi)}}{N} = \sqrt{\frac{\pi(1-\pi)}{N}} \tag{5.8.5}$$

Returning to the voter example, $\pi = 0.60$ and $N = 10$. (Don't confuse $\pi = 0.60$, the population proportion and $p = 0.70$, the sample proportion.) Therefore, the mean of the sampling distribution of p is 0.60. The standard error is

$$\sigma_p = \sqrt{\frac{0.60(1-0.60)}{10}} = 0.155 \tag{5.8.6}$$

The sampling distribution of p is a discrete rather than a continuous distribution. For example, with an N of 10, it is possible to have a p of 0.50 or a p of 0.60 but not a p of 0.55.

The sampling distribution of p is approximately normally distributed if N is fairly large and π is not close to 0 or 1. A rule of thumb is that the approximation is good if both $N\pi$ and $N(1-\pi)$ are greater than 10. The sampling distribution for the voter example is shown in Figure $5.8.1$. Note that even though $N(1-\pi)$ is only 4, the approximation is quite good.

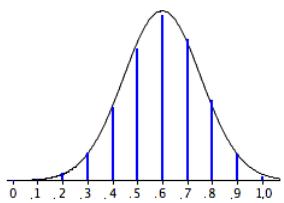


Figure $5.8.1$: The sampling distribution of p. Vertical bars are the probabilities; the smooth curve is the normal approximation

5.9: Statistical Literacy

Learning Objectives

- Accuracy of Employment Figures

The monthly jobs report always gets a lot of attention. Presidential candidates refer to the report when it favors their position. Referring to the August 2012 report in which only $96,000$ jobs were created, Republican presidential challenger Mitt Romney stated "the weak jobs report is devastating news for American workers and American families ... a harsh indictment of the president's handling of the economy." When the September 2012 report was released showing $114,000$ jobs were created (and the previous report was revised upwards), some supporters of Romney claimed the data were tampered with for political reasons. The most famous statement, "Unbelievable jobs numbers...these Chicago guys will do anything..can't debate so change numbers," was made by former Chairman and CEO of General Electric.

Example $5.9.1$: what do you think?

The standard error of the monthly estimate is $55,000$. Given that, what do you think of the difference between the two job reports?

Solution

The difference between the two reports is very small given that the standard error is $55,000$. It is not sensible to take any single jobs report too seriously.

5.E: Sampling Distributions (Exercises)

You may want to use the "r to z' calculator" and the "Calculate Area for a given X" applet for some of these exercises.

General Questions

Q1

A population has a mean of 50 and a standard deviation of 6.

a. What are the mean and standard deviation of the sampling distribution of the mean for $N = 16$?
b. What are the mean and standard deviation of the sampling distribution of the mean for $N = 20$? (relevant section)

Q2

Given a test that is normally distributed with a mean of 100 and a standard deviation of 12, find:

a. the probability that a single score drawn at random will be greater than 110 (relevant section)
b. the probability that a sample of 25 scores will have a mean greater than 105 (relevant section)
c. the probability that a sample of 64 scores will have a mean greater than 105 (relevant section)
d. the probability that the mean of a sample of 16 scores will be either less than 95 or greater than 105 (relevant section)

Q3

What term refers to the standard deviation of the sampling distribution? (relevant section)

Q4

a. If the standard error of the mean is 10 for $N = 12$, what is the standard error of the mean for $N = 22$?
b. If the standard error of the mean is 50 for $N = 25$, what is it for $N = 64$? (relevant section)

Q5

A questionnaire is developed to assess women's and men's attitudes toward using animals in research. One question asks whether animal research is wrong and is answered on a 7-point scale. Assume that in the population, the mean for women is 5, the mean for men is 4, and the standard deviation for both groups is 1.5. Assume the scores are normally distributed. If 12 women and 12 men are selected randomly, what is the probability that the mean of the women will be more than 1.5 points higher than the mean of the men? (relevant section)

Q6

If the correlation between reading achievement and math achievement in the population of fifth graders were 0.60, what would be the probability that in a sample of 28 students, the sample correlation coefficient would be greater than 0.65? (relevant section)

Q7

If numerous samples of $N = 15$ are taken from a uniform distribution and a relative frequency distribution of the means is drawn, what would be the shape of the frequency distribution? (relevant section & relevant section)

Q8

A normal distribution has a mean of 20 and a standard deviation of 10. Two scores are sampled randomly from the distribution and the second score is subtracted from the first. What is the probability that the difference score will be greater than 5? Hint: Read the Variance Sum Law section of Chapter 3. (relevant section & relevant section)

Q9

What is the shape of the sampling distribution of r? In what way does the shape depend on the size of the population correlation? (relevant section)

Q10

If you sample one number from a standard normal distribution, what is the probability it will be 0.5? (relevant section & relevant section)

Q11

A variable is normally distributed with a mean of 120 and a standard deviation of 5. Four scores are randomly sampled. What is the probability that the mean of the four scores is above 127? (relevant section)

Q12

The correlation between self esteem and extraversion is 0.30. A sample of 84 is taken.

a. What is the probability that the correlation will be less than 0.10?
b. What is the probability that the correlation will be greater than 0.25? (relevant section)

Q13

The mean GPA for students in School A is 3.0; the mean GPA for students in School B is 2.8. The standard deviation in both schools is 0.25. The GPAs of both schools are normally distributed. If 9 students are randomly sampled from each school, what is the probability that:

a. the sample mean for School A will exceed that of School A by 0.5 or more? (relevant section)
b. the sample mean for School B will be greater than the sample mean for School A? (relevant section)

Q14

In a city, 70% of the people prefer Candidate A. Suppose 30 people from this city were sampled.

a. What is the mean of the sampling distribution of p?
b. What is the standard error of p?
c. What is the probability that 80% or more of this sample will prefer Candidate A?
d. What is the probability that 45% or more of this sample will prefer some other candidate? (relevant section)

Q15

When solving problems where you need the sampling distribution of r, what is the reason for converting from r to z'? (relevant section)

Q16

In the population, the mean SAT score is 1000. Would you be more likely (or equally likely) to get a sample mean of 1200 if you randomly sampled 10 students or if you randomly sampled 30 students? Explain. (relevant section & relevant section)

Q17

True/false: The standard error of the mean is smaller when $N = 20$ than when $N = 10$. (relevant section)

Q18

True/false: The sampling distribution of $r = 0.8$ becomes normal as N increases. (relevant section)

Q19

True/false: You choose 20 students from the population and calculate the mean of their test scores. You repeat this process 100 times and plot the distribution of the means. In this case, the sample size is 100. (relevant section & relevant section)

Q20

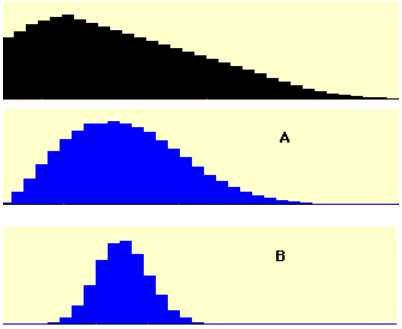
True/false: In your school, 40% of students watch TV at night. You randomly ask 5 students every day if they watch TV at night. Every day, you would find that 2 of the 5 do watch TV at night. (relevant section & relevant section)

Q21

True/false: The median has a sampling distribution. (relevant section)

Q22

True/false: Refer to the figure below. The population distribution is shown in black, and its corresponding sampling distribution of the mean for $N = 10$ is labeled "A" (relevant section & relevant section)



Questions from Case Studies

The following questions use data from the Angry Moods (AM) case study.

Q23

a. How many men were sampled?
b. How many women were sampled?

Q24

What is the mean difference between men and women on the Anger-Out scores?

Q25

Suppose in the population, the Anger-Out score for men is two points higher than it is for women. The population variances for men and women are both 20. Assume the Anger-Out scores for both genders are normally distributed. Given this information about the population parameters:

a. What is the mean of the sampling distribution of the difference between means? (relevant section)
b. What is the standard error of the difference between means? (relevant section)
c. What is the probability that you would have gotten this mean difference (see #24) or less in your sample? (relevant section)

The following questions use data from the Animal Research (AR) case study.

Q26

How many people were sampled to give their opinions on animal research?

Q27

(AR#11) What is the correlation in this sample between the belief that animal research is wrong and belief that animal research is necessary? (Ch. 4.E)

Q28

Suppose the correlation between the belief that animal research is wrong and the belief that animal research is necessary is -0.68 in the population.

a. Convert -0.68 to z'. (relevant section)
b. Find the standard error of this sampling distribution. (relevant section)
c. In a new sample, what is the probability that you would get the correlation found in the original sample (see #27) or a lower correlation (closer to -1)? (relevant section)

Selected Answers

S1

a. Mean = 50, SD = 1.5

S2

b. 0.019

S4

a. 7.39

S11

0.0026

S12

b. 0.690

S13

a. 0.0055

S14

c. 0.116

S23

a. 30

S25

a. 2

S28

c. 0.603

CHAPTER OVERVIEW

6: Logic of Hypothesis Testing

When interpreting an experimental finding, a natural question arises as to whether the finding could have occurred by chance. Hypothesis testing is a statistical procedure for testing whether chance is a plausible explanation of an experimental finding. Misconceptions about hypothesis testing are common among practitioners as well as students. To help prevent these misconceptions, this chapter goes into more detail about the logic of hypothesis testing than is typical for an introductory-level text.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

6.1: Introduction to Hypothesis Testing

Learning Objectives

- Describe the logic by which it can be concluded that someone can distinguish between two things
- State whether random assignment ensures that all uncontrolled sources of variation will be equal
- Define precisely what the probability is that is computed to reach the conclusion that a difference is not due to chance
- Distinguish between the probability of an event and the probability of a state of the world
- Define "null hypothesis"
- Be able to determine the null hypothesis from a description of an experiment
- Define "alternative hypothesis"

The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here we will present an example based on James Bond who insisted that martinis should be shaken rather than stirred. Let's consider a hypothetical experiment to determine whether Mr. Bond can tell the difference between a shaken and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Let's say Mr. Bond was correct on 13 of the 16 taste tests. Does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

This result does not prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct $13/16$ times or more. This probability can be computed from the binomial distribution, and the binomial distribution calculator shows it to be 0.0106. This is a pretty low probability, and therefore someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. So either Mr. Bond was very lucky, or he can tell whether the drink was shaken or stirred. The hypothesis that he was guessing is not proven false, but considerable doubt is cast on it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

Binomial Calculator

Let's consider another example. The case study Physicians' Reactions sought to determine whether physicians spend less time with obese patients. Physicians were sampled randomly and each was shown a chart of a patient complaining of a migraine headache. They were then asked to estimate how long they would spend with the patient. The charts were identical except that for half the charts, the patient was obese and for the other half, the patient was of average weight. The chart a particular physician viewed was determined randomly. Thirty-three physicians viewed charts of average-weight patients and 38 physicians viewed charts of obese patients.

The mean time physicians reported that they would spend with obese patients was 24.7 minutes as compared to a mean of 31.4 minutes for average-weight patients. How might this difference between means have occurred? One possibility is that physicians were influenced by the weight of the patients. On the other hand, perhaps by chance, the physicians who viewed charts of the obese patients tend to see patients for less time than the other physicians. Random assignmentof charts does not ensure that the groups will be equal in all respects other than the chart they viewed. In fact, it is certain the two groups differed in many ways by chance. The two groups could not have exactly the same mean age (if measured precisely enough such as in days). Perhaps a physician's age affects how long physicians see patients. There are innumerable differences between the groups that could affect how long they view patients. With this in mind, is it plausible that these chance differences are responsible for the difference in times?

To assess the plausibility of the hypothesis that the difference in mean times is due to chance, we compute the probability of getting a difference as large or larger than the observed difference ($31.4 - 24.7 = 6.7$ minutes) if the difference were, in fact, due solely to chance. Using methods presented in another section,this probability can be computed to be 0.0057. Since this is such a low probability, we have confidence that the difference in times is due to the patient's weight and is not due to chance.

The Probability Value

It is very important to understand precisely what the probability values mean. In the James Bond example, the computed probability of $0.0106$is the probability he would be correct on 13 or more taste tests (out of 16) if he were just guessing.

It is easy to mistake this probability of 0.0106as the probability he cannot tell the difference. This is not at all what it means.

The probability of 0.0106 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing). It is not the probability that a state of the world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim, the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that the probability of any number being evenly divisible by 7 is 0.50. The bird is correct on $9/16$ choices. Using the binomial calculator, we can compute that the probability of being correct nine or more times out of 16 if one is only guessing is 0.40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a 0.40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower than 0.0001.

To reiterate, the probability value is the probability of an outcome ($9/16$ or better) and not the probability of a particular state of the world (the bird was only guessing). In statistics, it is conventional to refer to possible states of the world as hypotheses since they are hypothesized states of the world. Using this terminology, the probability value is the probability of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis is false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he cannot tell the difference between shaken and stirred martinis. The probability value is low (0.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference. A branch of statistics called Bayesian statistics provides methods for computing the probabilities of hypotheses. These computations require that one specify the probability of the hypothesis before the data are considered and, therefore, are difficult to apply in some contexts.

The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the null hypothesis. In the Physicians' Reactions example, the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as:

$$\mu_{obese} = \mu_{average} \tag{6.1.1}$$

or as

$$\mu_{obese} - \mu_{average} = 0 \tag{6.1.2}$$

The null hypothesis in a correlational study of the relationship between high school grades and college grades would typically be that the population correlation is 0. This can be written as

$$\rho = 0 \tag{6.1.3}$$

where ρ is the population correlation (not to be confused with r, the correlation in the sample).

Although the null hypothesis is usually that the value of a parameter is 0, there are occasions in which the null hypothesis is a value other than 0. For example, if one were testing whether a subject differed from chance in their ability to determine whether a flipped coin would come up heads or tails, the null hypothesis would be that $\pi = 0.5$.

Keep in mind that the null hypothesis is typically the opposite of the researcher's hypothesis. In the Physicians' Reactions study, the researchers hypothesized that physicians would expect to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically is put forward with the hope that it can be discredited and therefore rejected. If the null hypothesis were true, a difference as large or larger than the sample difference of 6.7 minutes would be very unlikely to occur. Therefore, the researchers rejected the null hypothesis of no difference and concluded that in the population, physicians intend to spend less time with obese patients.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted. The alternative hypothesis is simply the reverse of the null hypothesis. If the null hypothesis $\mu_{obese} = \mu_{average}$ is rejected, then there are two alternatives:

$$\mu_{obese} < \mu_{average} \tag{6.1.4}$$

or

$$\mu_{obese} > \mu_{average} \tag{6.1.5}$$

Naturally, the direction of the sample means determines which alternative is adopted. Some textbooks have incorrectly argued that rejecting the null hypothesis that two population means are equal does not justify a conclusion about which population mean is larger. Kaiser (1960) showed how it is justified to draw a conclusion about the direction of the difference.

1. Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review*, *67*, 160-167.

6.2: Significance Testing

Learning Objectives

- Describe how a probability value is used to cast doubt on the null hypothesis
- Define "statistically significant"
- Distinguish between statistical significance and practical significance
- Distinguish between two approaches to significance testing

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01. When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the α (alpha) level or simply α. It is also called the significance level.

When the null hypothesis is rejected, the effect is said to be statistically significant. For example, in the Physicians' Reactions case study, the probability value is 0.0057. Therefore, the effect of obesity is statistically significant and the null hypothesis that obesity makes no difference is rejected. It is very important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what "significant" usually means. When an effect is significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is.

Do not confuse statistical significance with practical significance. A small effect can be highly significant if the sample size is large enough.

Why does the word "significant" in the phrase "statistically significant" mean something so different from other uses of the word? Interestingly, this is because the meaning of "significant" in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was "significant" if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is real and not due to chance. Over the years, the meaning of "significant" changed, leading to the potential misinterpretation.

There are two approaches (at least) to conducting significance tests. In one (favored by R. Fisher), a significance test is conducted and the probability value reflects the strength of the evidence against the null hypothesis. If the probability is below 0.01, the data provide strong evidence that the null hypothesis is false. If the probability value is below 0.05 but larger than 0.01, then the null hypothesis is typically rejected, but not with as much confidence as it would be if the probability value were below 0.01. Probability values between 0.05 and 0.10 provide weak evidence against the null hypothesis and, by convention, are not considered low enough to justify rejecting it. Higher probabilities provide less evidence that the null hypothesis is false.

The alternative approach (favored by the statisticians Neyman and Pearson) is to specify an α level before analyzing the data. If the data analysis results in a probability value below the α level, then the null hypothesis is rejected; if it is not, then the null hypothesis is not rejected. According to this perspective, if a result is significant, then it does not matter how significant it is. Moreover, if it is not significant, then it does not matter how close to being significant it is. Therefore, if the 0.05 level is being used, then probability values of 0.049 and 0.001 are treated identically. Similarly, probability values of 0.06 and 0.34 are treated identically.

The former approach (preferred by Fisher) is more suitable for scientific research and will be adopted here. The latter is more suitable for applications in which a yes/no decision must be made. For example, if a statistical analysis were undertaken to determine whether a machine in a manufacturing plant were malfunctioning, the statistical analysis would be used to determine whether or not the machine should be shut down for repair. The plant manager would be less interested in assessing the weight of the evidence than knowing what action should be taken. There is no need for an immediate decision in scientific research where a researcher may conclude that there is some evidence against the null hypothesis, but that more research is needed before a definitive conclusion can be drawn.

- **11.2: Significance Testing** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

6.3: Type I and II Errors

Learning Objectives

- Define **Type I** and **Type II** errors
- Interpret significant and non-significant differences
- Explain why the null hypothesis should not be accepted when the effect is not significant

In the Physicians' Reactions case study, the probability value associated with the significance test is 0.0057. Therefore, the null hypothesis was rejected, and it was concluded that physicians intend to spend less time with obese patients. Despite the low probability value, it is possible that the null hypothesis of no true difference between obese and average-weight patients is true and that the large difference between sample means occurred by chance. If this is the case, then the conclusion that physicians intend to spend less time with obese patients is in error. This type of error is called a **Type I** error. More generally, a **Type I** error occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05, then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the α (alpha) level or simply α. It is also called the significance level. As discussed in the section on significance testing, it is better to interpret the probability value as an indication of the weight of evidence against the null hypothesis than as part of a decision rule for making a reject or do-not-reject decision. Therefore, keep in mind that rejecting the null hypothesis is not an all-or-nothing decision.

The **Type I** error rate is affected by the α level: the lower the α level, the lower the **Type I** error rate. It might seem that α is the probability of a **Type I** error. However, this is not correct. Instead, α is the probability of a **Type I** error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a **Type I** error.

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a **Type II** error. Unlike a **Type I** error, a **Type II** error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a **Type I** error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true.

A **Type II** error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a **Type II** error is called β (beta). The probability of correctly rejecting a false null hypothesis equals $1-\beta$ and is called power. Power is covered in detail in another section.

6.4: One- and Two-Tailed Tests

Learning Objectives

- Define **Type I** and **Type II** errors
- Interpret significant and non-significant differences
- Explain why the null hypothesis should not be accepted when the effect is not significant

In the James Bond case study, Mr. Bond was given 16 trials on which he judged whether a martini had been shaken or stirred. He was correct on 13 of the trials. From the binomial distribution, we know that the probability of being correct 13 or more times out of 16 if one is only guessing is 0.0106. Figure $6.4.1$ shows a graph of the binomial distribution. The red bars show the values greater than or equal to 13. As you can see in the figure, the probabilities are calculated for the upper tail of the distribution. A probability calculated in only one tail of the distribution is called a "one-tailed probability."

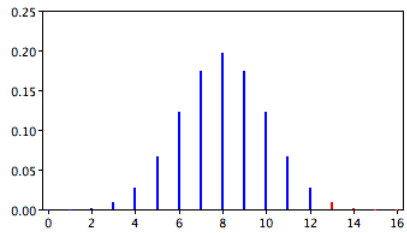


Figure $6.4.1$: *The binomial distribution. The upper (right-hand) tail is red*

Binomial Calculator

A slightly different question can be asked of the data: "What is the probability of getting a result as extreme or more extreme than the one observed?" Since the chance expectation is $8/16$, a result of $3/16$ is equally as extreme as $13/16$. Thus, to calculate this probability, we would consider both tails of the distribution. Since the binomial distribution is symmetric when $\pi = 0.5$, this probability is exactly double the probability of 0.0106 computed previously. Therefore, $p = 0.0212$. A probability calculated in both tails of a distribution is called a "two-tailed probability" (see Figure $6.4.2$).

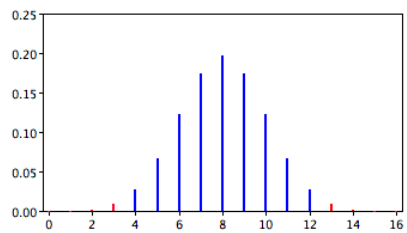


Figure $6.4.2$: *The binomial distribution. Both tails are red*

Should the one-tailed or the two-tailed probability be used to assess Mr. Bond's performance? That depends on the way the question is posed. If we are asking whether Mr. Bond can tell the difference between shaken or stirred martinis, then we would conclude he could if he performed either much better than chance or much worse than chance. If he performed much worse than chance, we would conclude that he can tell the difference, but he does not know which is which. Therefore, since we are going to reject the null hypothesis if Mr. Bond does either very well or very poorly, we will use a two-tailed probability.

On the other hand, if our question is whether Mr. Bond is better than chance at determining whether a martini is shaken or stirred, we would use a one-tailed probability. What would the one-tailed probability be if Mr. Bond were correct on only 3 of the 16 trials? Since the one-tailed probability is the probability of the right-hand tail, it would be the probability of getting 3 or more correct out of 16. This is a very high probability and the null hypothesis would not be rejected.

The null hypothesis for the two-tailed test is $\pi = 0.5$. By contrast, the null hypothesis for the one-tailed test is $\pi \leq 0.5$. Accordingly, we reject the two-tailed hypothesis if the sample proportion deviates greatly from 0.5 in either direction. The one-tailed hypothesis is rejected only if the sample proportion is much greater than 0.5. The alternative hypothesis in the two-tailed test is $\pi \neq 0.5$. In the one-tailed test it is $\pi > 0.5$.

You should always decide whether you are going to use a one-tailed or a two-tailed probability before looking at the data. Statistical tests that compute one-tailed probabilities are called one-tailed tests; those that compute two-tailed probabilities are

called two-tailed tests. Two-tailed tests are much more common than one-tailed tests in scientific research because an outcome signifying that something other than chance is operating is usually worth noting. One-tailed tests are appropriate when it is not important to distinguish between no effect and an effect in the unexpected direction. For example, consider an experiment designed to test the efficacy of a treatment for the common cold. The researcher would only be interested in whether the treatment was better than a placebo control. It would not be worth distinguishing between the case in which the treatment was worse than a placebo and the case in which it was the same because in both cases the drug would be worthless.

Some have argued that a one-tailed test is justified whenever the researcher predicts the direction of an effect. The problem with this argument is that if the effect comes out strongly in the non-predicted direction, the researcher is not justified in concluding that the effect is not zero. Since this is unrealistic, one-tailed tests are usually viewed skeptically if justified on this basis alone.

6.5: Significant Results

Learning Objectives

- Discuss whether rejection of the null hypothesis should be an all-or-none proposition
- State the usefulness of a significance test when it is extremely likely that the null hypothesis of no difference is false even before doing the experiment

When a probability value is below the α level, the effect is *statistically significant* and the null hypothesis is rejected. However, not all statistically significant effects should be treated the same way. For example, you should have less confidence that the null hypothesis is false if $p = 0.049$ than $p = 0.003$. Thus, rejecting the null hypothesis is not an all-or-none proposition.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted. Consider the one-tailed test in the James Bond case study: Mr. Bond was given 16 trials on which he judged whether a martini had been shaken or stirred and the question is whether he is better than chance on this task. The null hypothesis for this one-tailed test is that $\pi \leq 0.5$, where π is the probability of being correct on any given trial. If this null hypothesis is rejected, then the alternative hypothesis that $\pi > 0.5$ is accepted. If π is greater than 0.5, then Mr. Bond is better than chance on this task.

Now consider the two-tailed test used in the Physicians' Reactions case study. The null hypothesis is:

$$\mu_{obese} = \mu_{average} \tag{6.5.1}$$

If this null hypothesis is rejected, then there are two alternatives:

$$\mu_{obese} < \mu_{average} \tag{6.5.2}$$

$$\mu_{obese} > \mu_{average} \tag{6.5.3}$$

Naturally, the direction of the sample means determines which alternative is adopted. If the sample mean for the obese patients is significantly lower than the sample mean for the average-weight patients, then one should conclude that the population mean for the obese patients is lower than the population mean for the average-weight patients.

There are many situations in which it is very unlikely two conditions will have exactly the same population means. For example, it is practically impossible that aspirin and acetaminophen provide exactly the same degree of pain relief. Therefore, even before an experiment comparing their effectiveness is conducted, the researcher knows that the null hypothesis of exactly no difference is false. However, the researcher does not know which drug offers more relief. If a test of the difference is significant, then the direction of the difference is established. This point is also made in the section on the relationship between confidence intervals and significance tests.

Note

Some textbooks have incorrectly stated that rejecting the null hypothesis that two population means are equal does not justify a conclusion about which population mean is larger. Instead, they say that all one can conclude is that the population means differ. The validity of concluding the direction of the effect is clear if you note that a two-tailed test at the 0.05 level is equivalent to two separate one-tailed tests each at the 0.025 level. The two null hypotheses are then

$$\mu_{obese} \geq \mu_{average} \tag{6.5.4}$$

$$\mu_{obese} \leq \mu_{average} \tag{6.5.5}$$

If the former of these is rejected, then the conclusion is that the population mean for obese patients is lower than that for average-weight patients. If the latter is rejected, then the conclusion is that the population mean for obese patients is higher than that for average-weight patients.

6.6: Non-Significant Results

Learning Objectives

- Explain why the null hypothesis should not be accepted
- Discuss the problems of affirming a negative conclusion

When a significance test results in a high probability value, it means that the data provide little or no evidence that the null hypothesis is false. However, the high probability value is not evidence that the null hypothesis is true. The problem is that it is impossible to distinguish a null effect from a very small effect.

Example $6.6.1$

For example, in the James Bond Case Study, suppose Mr. Bond is, in fact, just barely better than chance at judging whether a martini was shaken or stirred. Assume he has a 0.51 probability of being correct on a given trial $\pi = 0.51$. Let's say Experimenter Jones (who did not know $\pi = 0.51$ tested Mr. Bond and found he was correct 49 times out of 100 tries. How would the significance test come out?

Solution

The experimenter's significance test would be based on the assumption that Mr. Bond has a 0.50 probability of being correct on each trial $\pi = 0.50$. Given this assumption, the probability of his being correct 49 or more times out of 100 is 0.62. This means that the probability value is 0.62, a value very much higher than the conventional significance level of 0.05. This result, therefore, does not give even a hint that the null hypothesis is false. However, we know (but Experimenter Jones does not) that $\pi = 0.51$ and not 0.50 and therefore that the null hypothesis is false. So, if Experimenter Jones had concluded that the null hypothesis was true based on the statistical analysis, he or she would have been mistaken.

Concluding that the null hypothesis is true is called accepting the null hypothesis. To do so is a serious error.

Binomial Calculator

Further argument for not accepting the null hypothesis

Do not accept the null hypothesis when you do not reject it.

So how should the non-significant result be interpreted? The experimenter should report that there is no credible evidence Mr. Bond can tell whether a martini was shaken or stirred, but that there is no proof that he cannot. It is generally impossible to prove a negative. What if I claimed to have been Socrates in an earlier life? Since I have no evidence for this claim, I would have great difficulty convincing anyone that it is true. However, no one would be able to prove definitively that I was not.

Often a non-significant finding increases one's confidence that the null hypothesis is false. Consider the following hypothetical example.

Example $6.6.2$

A researcher develops a treatment for anxiety that he or she believes is better than the traditional treatment. A study is conducted to test the relative effectiveness of the two treatments: 20 subjects are randomly divided into two groups of 10. One group receives the new treatment and the other receives the traditional treatment. The mean anxiety level is lower for those receiving the new treatment than for those receiving the traditional

treatment. However, the difference is not significant. The statistical analysis shows that a difference as large or larger than the one obtained in the experiment would occur 11% of the time even if there were no true difference between the treatments. In other words, the probability value is 0.11. A naive researcher would interpret this finding as evidence that the new treatment is no more effective than the traditional treatment. However, the sophisticated researcher, although disappointed that the effect was not significant, would be encouraged that the new treatment led to less anxiety than the traditional treatment. The data support the thesis that the new treatment is better than the traditional one even though the effect is not statistically significant. This researcher should have more confidence that the new treatment is better than he or she had before the experiment was conducted. However, the support is weak and the data are inconclusive. What should the researcher do?

Solution

A reasonable course of action would be to do the experiment again. Let's say the researcher repeated the experiment and again found the new treatment was better than the traditional treatment. However, once again the effect was not significant and this time the probability value was 0.07. The naive researcher would think that two out of two experiments failed to find significance and therefore the new treatment is unlikely to be better than the traditional treatment. The sophisticated researcher would note that two out of two times the new treatment was better than the traditional treatment. Moreover, two experiments each providing weak support that the new treatment is better, when taken together, can provide strong support. Using a method for combining probabilities, it can be determined that combining the probability values of 0.11 and 0.07 results in a probability value of 0.045. Therefore, these two non-significant findings taken together result in a significant finding.

Although there is never a statistical basis for concluding that an effect is exactly zero, a statistical analysis can demonstrate that an effect is most likely small. This is done by computing a confidence interval. If all effect sizes in the interval are small, then it can be concluded that the effect is small. For example, suppose an experiment tested the effectiveness of a treatment for insomnia. Assume that the mean time to fall asleep was 2 minutes shorter for those receiving the treatment than for those in the control group and that this difference was not significant. If the 95% confidence interval ranged from -4 to 8 minutes, then the researcher would be justified in concluding that the benefit is eight minutes or less. However, the researcher would not be justified in concluding the null hypothesis is true, or even that it was supported.

6.7: Steps in Hypothesis Testing

Learning Objectives

- Be able to state the null hypothesis for both one-tailed and two-tailed tests
- Differentiate between a significance level and a probability level
- State the four steps involved in significance testing

1. The first step is to specify the null hypothesis. For a two-tailed test, the null hypothesis is typically that a parameter equals zero although there are exceptions. A typical null hypothesis is $\mu_1 - \mu_2 = 0$ which is equivalent to $\mu_1 = \mu_2$. For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero. If the prediction is that μ_1 is larger than μ_2, then the null hypothesis (the reverse of the prediction) is $\mu_1 - \mu_2 \geq 0$. This is equivalent to $\mu_1 \leq \mu_2$.
2. The second step is to specify the α level which is also known as the significance level. Typical values are 0.05 and 0.01.
3. The third step is to compute the probability value (also known as the p value). This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.
4. Finally, compare the probability value with the α level. If the probability value is lower then you reject the null hypothesis. Keep in mind that rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false. However, if your probability value is higher than the conventional α level of 0.05, most scientists will consider your findings inconclusive. Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it.

6.8: Significance Testing and Confidence Intervals

Learning Objectives

- Explain why a confidence interval makes clear that one should not accept the null hypothesis

There is a close relationship between confidence intervals and significance tests. Specifically, if a statistic is significantly different from 0 at the 0.05 level, then the 95% confidence interval will not contain 0. All values in the confidence interval are plausible values for the parameter, whereas values outside the interval are rejected as plausible values for the parameter. In the Physicians' Reactions case study, the 95% confidence interval for the difference between means extends from 2.00 to 11.26. Therefore, any value lower than 2.00 or higher than 11.26 is rejected as a plausible value for the population difference between means. Since zero is lower than 2.00, it is rejected as a plausible value and a test of the null hypothesis that there is no difference between means is significant. It turns out that the p value is 0.0057. There is a similar relationship between the 99% confidence interval and significance at the 0.01 level.

Whenever an effect is significant, all values in the confidence interval will be on the same side of zero (either all positive or all negative). Therefore, a significant finding allows the researcher to specify the direction of the effect. There are many situations in which it is very unlikely two conditions will have exactly the same population means. For example, it is practically impossible that aspirin and acetaminophen provide exactly the same degree of pain relief. Therefore, even before an experiment comparing their effectiveness is conducted, the researcher knows that the null hypothesis of exactly no difference is false. However, the researcher does not know which drug offers more relief. If a test of the difference is significant, then the direction of the difference is established because the values in the confidence interval are either all positive or all negative.

If the 95% confidence interval contains zero (more precisely, the parameter value specified in the null hypothesis), then the effect will not be significant at the 0.05 level. Looking at non-significant effects in terms of confidence intervals makes clear why the null hypothesis should not be accepted when it is not rejected: Every value in the confidence interval is a plausible value of the parameter. Since zero is in the interval, it cannot be rejected. However, there is an infinite number of other values in the interval (assuming continuous measurement), and none of them can be rejected either.

6.9: Misconceptions of Hypothesis Testing

Learning Objectives

- State why the probability value is not the probability the null hypothesis is false
- Explain why a low probability value does not necessarily mean there is a large effect
- Explain why a non-significant outcome does not mean the null hypothesis is probably true

Misconceptions about significance testing are common. This section lists three important ones.

1. Misconception: The probability value is the probability that the null hypothesis is false.

   Proper interpretation: The probability value is the probability of a result as extreme or more extreme given that the null hypothesis is true. It is the probability of the data given the null hypothesis. It is not the probability that the null hypothesis is false.
2. Misconception: A low probability value indicates a large effect.

   Proper interpretation: A low probability value indicates that the sample outcome (or one more extreme) would be very unlikely if the null hypothesis were true. A low probability value can occur with small effect sizes, particularly if the sample size is large.
3. Misconception: A non-significant outcome means that the null hypothesis is probably true.

   Proper interpretation: A non-significant outcome means that the data do not conclusively demonstrate that the null hypothesis is false.

6.10: Statistical Literacy

Learning Objectives

- Evidence for the Higgs Boson

Research in March, 2012 reported here found evidence for the existence of the Higgs Boson particle. However, the evidence for the existence of the particle was not statistically significant.

Example 6.10.1:what do you think?

Did the researchers conclude that their investigation had been a failure or did they conclude they have evidence of the particle, just not strong enough evidence to draw a confident conclusion?

Solution

One of the investigators stated, "We see some tantalizing evidence but not significant enough to make a stronger statement." Therefore, they were encouraged by the result. In a subsequent study, the evidence was significant.

6.E: Logic of Hypothesis Testing (Exercises)

You may want to use the Binomial Calculator for some of these exercises.

General Questions

Q1

An experiment is conducted to test the claim that James Bond can taste the difference between a Martini that is shaken and one that is stirred. What is the null hypothesis? (relevant section)

Q2

The following explanation is incorrect. What three words should be added to make it correct? (relevant section)

> The probability value is the probability of obtaining a statistic as different from the parameter specified in the null hypothesis as the statistic obtained in the experiment. The probability value is computed assuming that the null hypothesis is true.

Q3

Why do experimenters test hypotheses they think are false? (relevant section)

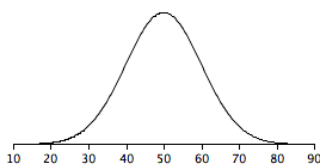
Q4

State the null hypothesis for:

a. An experiment testing whether echinacea decreases the length of colds.
b. A correlational study on the relationship between brain size and intelligence.
c. An investigation of whether a self-proclaimed psychic can predict the outcome of a coin flip.
d. A study comparing a drug with a placebo on the amount of pain relief. (A one-tailed test was used.)
(relevant section & relevant section)

Q5

Assume the null hypothesis is that $\mu = 50$ and that the graph shown below is the sampling distribution of the mean (M). Would a sample value of $M = 60$ be significant in a two-tailed test at the 0.05 level? Roughly what value of M would be needed to be significant? (relevant section & relevant section)



Q6

A researcher develops a new theory that predicts that vegetarians will have more of a particular vitamin in their blood than non-vegetarians. An experiment is conducted and vegetarians do have more of the vitamin, but the difference is not significant. The probability value is 0.13. Should the experimenter's confidence in the theory increase, decrease, or stay the same? (relevant section)

Q7

A researcher hypothesizes that the lowering in cholesterol associated with weight loss is really due to exercise. To test this, the researcher carefully controls for exercise while comparing the cholesterol levels of a group of subjects who lose weight by dieting with a control group that does not diet. The difference between groups in cholesterol is not significant. Can the researcher claim that weight loss has no effect? (relevant section)

Q8

A significance test is performed and $p = 0.20$. Why can't the experimenter claim that the probability that the null hypothesis is true is 0.20? (relevant section, relevant section & relevant section)

Q9

For a drug to be approved by the FDA, the drug must be shown to be safe and effective. If the drug is significantly more effective than a placebo, then the drug is deemed effective. What do you know about the effectiveness of a drug once it has been approved by the FDA (assuming that there has not been a **Type I** error)? (relevant section)

Q10

When is it valid to use a one-tailed test? What is the advantage of a one-tailed test? Give an example of a null hypothesis that would be tested by a one-tailed test. (relevant section)

Q11

Distinguish between probability value and significance level. (relevant section)

Q12

Suppose a study was conducted on the effectiveness of a class on "How to take tests." The SAT scores of an experimental group and a control group were compared. (There were 100 subjects in each group.) The mean score of the experimental group was 503 and the mean score of the control group was 499. The difference between means was found to be significant, $p = 0.037$. What do you conclude about the effectiveness of the class? (relevant section & relevant section)

Q13

Is it more conservative to use an alpha level of 0.01 or an alpha level of 0.05? Would beta be higher for an alpha of 0.05 or for an alpha of 0.01? (relevant section)

Q14

Why is $H_o : M_1 = M_2$ not a proper null hypothesis? (relevant section)

Q15

An experimenter expects an effect to come out in a certain direction. Is this sufficient basis for using a one-tailed test? Why or why not? (relevant section)

Q16

How do the **Type I** and **Type II** error rates of one-tailed and two-tailed tests differ? (relevant section & relevant section)

Q17

A two-tailed probability is 0.03. What is the one-tailed probability if the effect were in the specified direction? What would it be if the effect were in the other direction? (relevant section)

Q18

You choose an alpha level of 0.01 and then analyze your data.

a. What is the probability that you will make a **Type I** error given that the null hypothesis is true?
b. What is the probability that you will make a **Type I** error given that the null hypothesis is false? (relevant section)

Q19

Why doesn't it make sense to test the hypothesis that the sample mean is 42? (relevant section & relevant section)

Q20

True/false: It is easier to reject the null hypothesis if the researcher uses a smaller alpha (α) level. (relevant section & relevant section)

Q21

True/false: You are more likely to make a Type I error when using a small sample than when using a large sample. (relevant section)

Q22

True/false: You accept the alternative hypothesis when you reject the null hypothesis. (relevant section)

Q23

True/false: You do not accept the null hypothesis when you fail to reject it. (relevant section)

Q24

True/false: A researcher risks making a **Type I** error any time the null hypothesis is rejected. (relevant section)

CHAPTER OVERVIEW

7: Estimation

One of the major applications of statistics is estimating population parameters from sample statistics. For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium. Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 of the people supported the proposition. This value of 0.53 is called a point estimate of the population proportion. It is called a point estimate because the estimate consists of a single value or point

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

7.1: Introduction to Estimation

Learning Objectives

- Define statistic
- Define parameter
- Define point estimate
- Define interval estimate
- Define margin of error

One of the major applications of statistics is estimating population parameters from sample statistics. For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium. Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 of the people supported the proposition. This value of 0.53 is called a point estimate of the population proportion. It is called a point estimate because the estimate consists of a single value or point.

Point estimates are usually supplemented by interval estimates called confidence intervals. Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time. For example, if the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval: $0.46 < \pi < 0.60$. The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal. The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%.

In an experiment on memory for chess positions, the mean recall for tournament players was 63.8 and the mean for non-players was 33.1. Therefore a point estimate of the difference between population means is 30.7. The 95% confidence interval on the difference between means extends from 19.05 to 42.35. You will see how to compute this kind of interval in another section.

7.2: Degrees of Freedom

Learning Objectives

- Define degrees of freedom
- Estimate the variance from a sample of 1 if the population mean is known
- State why deviations from the sample mean are not independent
- State the general formula for degrees of freedom in terms of the number of values and the number of estimated parameters
- Calculate s^2

Some estimates are based on more information than others. For example, an estimate of the variance based on a sample size of 100 is based on more information than an estimate of the variance based on a sample size of 5. The degrees of freedom (df) of an estimate is the number of independent pieces of information on which the estimate is based.

As an example, let's say that we know that the mean height of Martians is 6 and wish to estimate the variance of their heights. We randomly sample one Martian and find that its height is 8. Recall that the variance is defined as the mean squared deviation of the values from their population mean. We can compute the squared deviation of our value of 8 from the population mean of 6 to find a single squared deviation from the mean. This single squared deviation from the mean, $(8-6)^2=4$, is an estimate of the mean squared deviation for all Martians. Therefore, based on this sample of one, we would estimate that the population variance is 4. This estimate is based on a single piece of information and therefore has 1 df. If we sampled another Martian and obtained a height of 5, then we could compute a second estimate of the variance, $(5-6)^2=1$. We could then average our two estimates (4 and 1) to obtain an estimate of 2.5. Since this estimate is based on two independent pieces of information, it has two degrees of freedom. The two estimates are independent because they are based on two independently and randomly selected Martians. The estimates would not be independent if after sampling one Martian, we decided to choose its brother as our second Martian.

As you are probably thinking, it is pretty rare that we know the population mean when we are estimating the variance. Instead, we have to first estimate the population mean (μ) with the sample mean (M). The process of estimating the mean affects our degrees of freedom as shown below.

Returning to our problem of estimating the variance in Martian heights, let's assume we do not know the population mean and therefore we have to estimate it from the sample. We have sampled two Martians and found that their heights are 8 and 5. Therefore M, our estimate of the population mean, is

$$M=\frac{(8+5)}{2}=6.5 \tag{7.2.1}$$

We can now compute two estimates of variance:

- Estimate $1=(8-6.5)^2=2.25$
- Estimate $2=(5-6.5)^2=2.25$

Now for the key question: Are these two estimates independent? The answer is no because each height contributed to the calculation of M. Since the first Martian's height of 8 influenced M, it also influenced Estimate 2. If the first height had been, for example, 10, then M would have been 7.5 and Estimate 2 would have been $(5-7.5)^2=6.25$ instead of 2.25. The important point is that the two estimates are not independent and therefore we do not have two degrees of freedom. Another way to think about the non-independence is to consider that if you knew the mean and one of the scores, you would know the other score. For example, if one score is 5 and the mean is 6.5, you can compute that the total of the two scores is 13 and therefore that the other score must be $13-5=8$.

In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. In the Martians example, there are two values (8 and 5) and we had to estimate one parameter (μ) on the way to estimating the parameter of interest (σ^2). Therefore, the estimate of variance has $2-1=1$ degree of freedom. If we had sampled 12 Martians, then our estimate of variance would have had 11 degrees of freedom. Therefore, the degrees of freedom of an estimate of variance is equal to $N-1$, where N is the number of observations.

Recall from the section on variability that the formula for estimating the variance in a sample is:

$$s^2=\frac{\sum(X-M)^2}{N-1} \tag{7.2.2}$$

The denominator of this formula is the degrees of freedom.

7.3: Characteristics of Estimators

Learning Objectives

1. Define bias
2. Define sampling variability
3. Define expected value
4. Define relative efficiency

This section discusses two important characteristics of statistics used as point estimates of parameters: bias and sampling variability. Bias refers to whether an estimator tends to either over or underestimate the parameter. Sampling variability refers to how much the estimate varies from sample to sample.

Have you ever noticed that some bathroom scales give you very different weights each time you weigh yourself? With this in mind, let's compare two scales. Scale 1 is a very high-tech digital scale and gives essentially the same weight each time you weigh yourself; it varies by at most 0.02 pounds from weighing to weighing. Although this scale has the potential to be very accurate, it is calibrated incorrectly and, on average, overstates your weight by one pound. Scale 2 is a cheap scale and gives very different results from weighing to weighing. However, it is just as likely to underestimate as overestimate your weight. Sometimes it vastly overestimates it and sometimes it vastly underestimates it. However, the average of a large number of measurements would be your actual weight. Scale 1 is biased since, on average, its measurements are one pound higher than your actual weight. Scale 2, by contrast, gives unbiased estimates of your weight. However, Scale 2 is highly variable and its measurements are often very far from your true weight. Scale 1, in spite of being biased, is fairly accurate. Its measurements are never more than 1.02 pounds from your actual weight.

We now turn to more formal definitions of variability and precision. However, the basic ideas are the same as in the bathroom scale example.

Bias

A statistic is biased if the long-term average value of the statistic is not the parameter it is estimating. More formally, a statistic is biased if the mean of the sampling distribution of the statistic is not equal to the parameter. The mean of the sampling distribution of a statistic is sometimes referred to as the expected value of the statistic.

As we saw in the section on the sampling distribution of the mean, the mean of the sampling distribution of the (sample) mean is the population mean (μ). Therefore the sample mean is an unbiased estimate of μ. Any given sample mean may underestimate or overestimate μ, but there is no systematic tendency for sample means to either under or overestimate μ.

In the section on variability, we saw that the formula for the variance in a population is

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \tag{7.3.1}$$

whereas the formula to estimate the variance from a sample is

$$s^2 = \frac{\sum(X-M)^2}{N-1} \tag{7.3.2}$$

Notice that the denominators of the formulas are different: N for the population and $N-1$ for the sample. We saw in the "Estimating Variance Simulation" that if N is used in the formula for s^2, then the estimates tend to be too low and therefore biased. The formula with $N-1$ in the denominator gives an unbiased estimate of the population variance. Note that $N-1$ is the degrees of freedom.

Sampling Variability

The sampling variability of a statistic refers to how much the statistic varies from sample to sample and is usually measured by its standard error ; the smaller the standard error, the less the sampling variability. For example, the standard error of the mean is a measure of the sampling variability of the mean. Recall that the formula for the standard error of the mean is

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{7.3.3}$$

The larger the sample size (N), the smaller the standard error of the mean and therefore the lower the sampling variability.

Statistics differ in their sampling variability even with the same sample size. For example, for normal distributions, the standard error of the median is larger than the standard error of the mean. The smaller the standard error of a statistic, the more efficient the statistic. The relative efficiency of two statistics is typically defined as the ratio of their standard errors. However, it is sometimes defined as the ratio of their squared standard errors.

7.4: Bias and Variability Simulation

Learning Objectives

- To explore various aspects of sampling distributions

This simulation lets you explore various aspects of sampling distributions. When it begins, a histogram of a normal distribution is displayed at the topic of the screen.

The distribution portrayed at the top of the screen is the population from which samples are taken. The mean of the distribution is indicated by a small blue line and the median is indicated by a small purple line. Since the mean and median are the same, the two lines overlap. The red line extends from the mean one standard deviation in each direction. Note the correspondence between the colors used on the histogram and the statistics displayed to the left of the histogram.

The second histogram displays the sample data. This histogram is initially blank. The third and fourth histograms show the distribution of statistics computed from the sample data. The number of samples (replications) that the third and fourth histograms are based on is indicated by the label "Reps=."

Basic operations

The simulation is set to initially sample five numbers from the population, compute the mean of the five numbers, and plot the mean. Click the "Animated sample" button and you will see the five numbers appear in the histogram. The mean of the five numbers will be computed and the mean will be plotted in the third histogram. Do this several times to see the distribution of means begin to be formed. Once you see how this works, you can speed things up by taking 5, $1,000$, or $10,000$ samples at a time.

Choosing a statistic

The following statistics can be computed from the samples by choosing form the pop-up menu:

1. Mean: Mean
2. SD: Standard deviation of the sample (N is used in the denominator)
3. Variance: Variance of the sample (N is used in the denominator)
4. Variance (U): Unbiased estimate of variance ($N-1$ is used in denominator)
5. MAD: Mean absolute value of the deviation from the mean
6. Range: Range

Selecting a sample size

The size of each sample can be set to $2, 5, 10, 16, 20$ or 25 from the pop-up menu. Be sure not to confuse sample size with number of samples.

Comparison to a normal distribution

By clicking the "Fit normal" button you can see a normal distribution superimposed over the simulated sampling distribution.

Changing the population distribution

You can change the population by clicking on the top histogram with the mouse and dragging

7.5: Confidence Intervals

Learning Objectives

- To learn how to compute confidence intervals for various parameters

The next few sections show how to compute confidence intervals for a variety of parameters listed below.

- Confidence Interval for the Mean
- t distribution
- Confidence Interval Simulation
- Confidence Interval for the Difference Between Means
- Confidence Interval for Pearson's Correlation
- Confidence Interval for a Proportion

The "Confidence Interval Simulation" allows you to explore various properties of confidence intervals.

7.6: Confidence Intervals Intro

Learning Objectives

- Define confidence interval
- State why a confidence interval is not the probability the interval contains the parameter

Say you were interested in the mean weight of 10-year-old girls living in the United States. Since it would have been impractical to weigh all the 10-year-old girls in the United States, you took a sample of 16 and found that the mean weight was 90 pounds. This sample mean of 90 is a point estimate of the population mean. A point estimate by itself is of limited usefulness because it does not reveal the uncertainty associated with the estimate; you do not have a good sense of how far this sample mean may be from the population mean. For example, can you be confident that the population mean is within 5 pounds of 90? You simply do not know.

Confidence intervals provide more information than point estimates. Confidence intervals for means are intervals constructed using a procedure (presented in the next section) that will contain the population mean a specified proportion of the time, typically either 95% or 99% of the time. These intervals are referred to as 95% and 99% confidence intervals respectively. An example of a 95% confidence interval is shown below:

$$72.85 < \mu < 107.15 \tag{7.6.1}$$

There is good reason to believe that the population mean lies between these two bounds of 72.85 and 107.15 since 95% of the time confidence intervals contain the true mean.

If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean. Naturally, 5% of the intervals would not contain the population mean.

It is natural to interpret a 95% confidence interval as an interval with a 0.95 probability of containing the population mean. However, the proper interpretation is not that simple. One problem is that the computation of a confidence interval does not take into account any other information you might have about the value of the population mean. For example, if numerous prior studies had all found sample means above 110, it would not make sense to conclude that there is a 0.95 probability that the population mean is between 72.85 and 107.15. What about situations in which there is no prior information about the value of the population mean? Even here the interpretation is complex. The problem is that there can be more than one procedure that produces intervals that contain the population parameter 95% of the time. Which procedure produces the "true" 95% confidence interval? Although the various methods are equal from a purely mathematical point of view, the standard method of computing confidence intervals has two desirable properties: each interval is symmetric about the point estimate and each interval is contiguous. Recall from the introductory section in the chapter on probability that, for some purposes, probability is best thought of as subjective. It is reasonable, although not required by the laws of probability, that one adopt a subjective probability of 0.95 that a 95% confidence interval, as typically computed, contains the parameter in question.

Confidence intervals can be computed for various parameters, not just the mean. For example, later in this chapter you will see how to compute a confidence interval for ρ, the population value of Pearson's r, based on sample data.

7.7: Confidence Interval for Mean

Learning Objectives

- Use the inverse normal distribution calculator to find the value of z to use for a confidence interval
- Compute a confidence interval on the mean when σ is known
- Determine whether to use a t distribution or a normal distribution
- Compute a confidence interval on the mean when σ is estimated

When you compute a confidence interval on the mean, you compute the mean of a sample in order to estimate the mean of the population. Clearly, if you already knew the population mean, there would be no need for a confidence interval. However, to explain how confidence intervals are constructed, we are going to work backwards and begin by assuming characteristics of the population. Then we will show how sample data can be used to construct a confidence interval.

Assume that the weights of 10-year-old children are normally distributed with a mean of 90 and a standard deviation of 36. What is the sampling distribution of the mean for a sample size of 9? Recall from the section on the sampling distribution of the mean that the mean of the sampling distribution is μ and the standard error of the mean is

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{7.7.1}$$

For the present example, the sampling distribution of the mean has a mean of 90 and a standard deviation of $36/3 = 12$. Note that the standard deviation of a sampling distribution is its standard error. Figure $7.7.1$ shows this distribution. The shaded area represents the middle 95% of the distribution and stretches from 66.48 to 113.52 These limits were computed by adding and subtracting 1.96 standard deviations to/from the mean of 90 as follows:

$$90 - (1.96)(12) = 66.48 \tag{7.7.2}$$

$$90 + (1.96)(12) = 113.52 \tag{7.7.3}$$

The value of 1.96 is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean; 12 is the standard error of the mean.

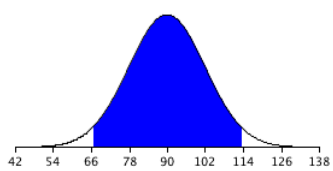


Figure $7.7.1$: The sampling distribution of the mean for $N = 9$. The middle 95% of the distribution is shaded

Figure $7.7.1$ shows that 95% of the means are no more than 23.52 units (1.96 standard deviations) from the mean of 90. Now consider the probability that a sample mean computed in a random sample is within 23.52 units of the population mean of 90. Since 95% of the distribution is within 23.52 of 90, the probability that the mean from any given sample will be within 23.52 of 90 is 0.95. This means that if we repeatedly compute the mean (M) from a sample, and create an interval ranging from $M - 23.52$ to $M + 23.52$, this interval will contain the population mean 95% of the time. In general, you compute the 95% confidence interval for the mean with the following formula:

$$\text{Lower limit} = M - Z_{0.95}\sigma_M \tag{7.7.4}$$

$$\text{Upper limit} = M + Z_{0.95}\sigma_M \tag{7.7.5}$$

where $Z_{0.95}$ is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area and σ_M is the standard error of the mean.

If you look closely at this formula for a confidence interval, you will notice that you need to know the standard deviation (σ) in order to estimate the mean. This may sound unrealistic, and it is. However, computing a confidence interval when σ is known is easier than when σ has to be estimated, and serves a pedagogical purpose. Later in this section we will show how to compute a confidence interval for the mean when σ has to be estimated.

Suppose the following five numbers were sampled from a normal distribution with a standard deviation of $2.5: 2, 3, 5, 6,\ and\ 9$. To compute the 95% confidence interval, start by computing the mean and standard error:

$$M = \frac{2+3+5+6+9}{5} = 5 \tag{7.7.6}$$

$$\sigma_M = \frac{2.5}{\sqrt{5}} = 1.118 \tag{7.7.7}$$

$Z_{0.95}$ can be found using the normal distribution calculator and specifying that the shaded area is 0.95 and indicating that you want the area to be between the cutoff points. As shown in Figure $7.7.2$, the value is 1.96. If you had wanted to compute the 99% confidence interval, you would have set the shaded area to 0.99 and the result would have been 2.58.

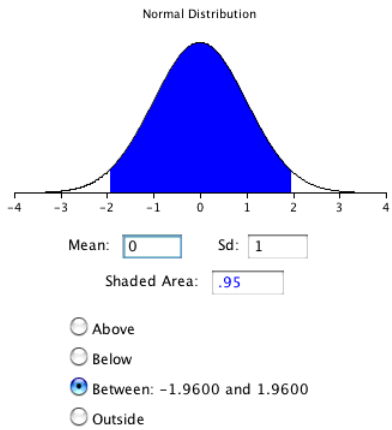


Figure $7.7.2$: 95% of the area is between -1.96 and 1.96

The confidence interval can then be computed as follows:

$$\text{Lower limit} = 5 - (1.96)(1.118) = 2.81 \tag{7.7.8}$$

$$\text{Upper limit} = 5 + (1.96)(1.118) = 7.19 \tag{7.7.9}$$

You should use the t distribution rather than the normal distribution when the variance is not known and has to be estimated from sample data. When the sample size is large, say 100 or above, the t distribution is very similar to the standard normal distribution. However, with smaller sample sizes, the t distribution is leptokurtic, which means it has relatively more scores in its tails than does the normal distribution. As a result, you have to extend farther from the mean to contain a given proportion of the area. Recall that with a normal distribution, 95% of the distribution is within 1.96 standard deviations of the mean. Using the t distribution, if you have a sample size of only 5, 95% of the area is within 2.78 standard deviations of the mean. Therefore, the standard error of the mean would be multiplied by 2.78 rather than 1.96.

The values of t to be used in a confidence interval can be looked up in a table of the t distribution. A small version of such a table is shown in Table $7.7.1$. The first column, df, stands for degrees of freedom, and for confidence intervals on the mean, df is equal to $N-1$, where N is the sample size.

Table $7.7.1$: Abbreviated t table

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032

8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

You can also use the "inverse t distribution" calculator to find the t values to use in confidence intervals. You will learn more about the t distribution in the next section.

Assume that the following five numbers are sampled from a normal distribution: $2, 3, 5, 6,\ and\ 9$ and that the standard deviation is not known. The first steps are to compute the sample mean and variance:

$$M = 5 \text{ and } S^2 = 7.5 \tag{7.7.10}$$

The next step is to estimate the standard error of the mean. If we knew the population variance, we could use the following formula:

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{7.7.11}$$

Instead we compute an estimate of the standard error (s_M):

$$s_M = \frac{s}{\sqrt{N}} = 1.225 \tag{7.7.12}$$

The next step is to find the value of t. As you can see from Table $7.7.1$, the value for the 95% interval for $df = N - 1 = 4$ is 2.776. The confidence interval is then computed just as it is when σ_M. The only differences are that s_M and t rather than σ_M and Z are used.

$$\text{Lower limit} = 5 - (2.776)(1.225) = 1.60 \tag{7.7.13}$$

$$\text{Upper limit} = 5 + (2.776)(1.225) = 8.40 \tag{7.7.14}$$

More generally, the formula for the 95% confidence interval on the mean is:

$$\text{Lower limit} = M - (t_{CL})(s_M) \tag{7.7.15}$$

$$\text{Upper limit} = M + (t_{CL})(s_M) \tag{7.7.16}$$

where M is the sample mean, t_{CL} is the t for the confidence level desired (0.95 in the above example), and s_M is the estimated standard error of the mean.

We will finish with an analysis of the Stroop Data. Specifically, we will compute a confidence interval on the mean difference score. Recall that 47 subjects named the color of ink that words were written in. The names conflicted so that, for example, they would name the ink color of the word "blue" written in red ink. The correct response is to say "red" and ignore the fact that the word is "blue." In a second condition, subjects named the ink color of colored rectangles.

Table $7.7.2$: Response times in seconds for 10 subjects

Naming Colored Rectangle	Interference	Difference
17	38	21
15	58	43
18	35	17

20	39	19
18	33	15
20	32	12
20	45	25
19	52	33
17	31	14
21	29	8

Table $7.7.2$ shows the time difference between the interference and color-naming conditions for 10 of the 47 subjects. The mean time difference for all 47 subjects is 16.362 seconds and the standard deviation is 7.470 seconds. The standard error of the mean is 1.090. A t table shows the critical value of t for $47-1=46$ degrees of freedom is 2.013 (for a 95% confidence interval). Therefore the confidence interval is computed as follows:

$$\text{Lower limit} = 16.362 - (2.013)(1.090) = 14.17 \tag{7.7.17}$$

$$\text{Upper limit} = 16.362 - (2.013)(1.090) = 18.56 \tag{7.7.18}$$

Therefore, the interference effect (difference) for the whole population is likely to be between 14.17 and 18.56 seconds.

7.8: t Distribution

Learning Objectives

- State the difference between the shape of the t distribution and the normal distribution
- State how the difference between the shape of the t distribution and normal distribution is affected by the degrees of freedom
- Use a t table to find the value of t to use in a confidence interval
- Use the t calculator to find the value of t to use in a confidence interval

In the introduction to normal distributions it was shown that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean. Therefore, if you randomly sampled a value from a normal distribution with a mean of 100, the probability it would be within 1.96σ of 100 is 0.95. Similarly, if you sample N values from the population, the probability that the sample mean (M) will be within $1.96\sigma_M$ of 100 is 0.95.

Now consider the case in which you have a normal distribution but you do not know the standard deviation. You sample N values and compute the sample mean (M) and estimate the standard error of the mean (σ_M) with s_M. What is the probability that M will be within $1.96s_M$ of the population mean (μ)?

This is a difficult problem because there are two ways in which M could be more than $1.96s_M$ from μ:

1. M could, by chance, be either very high or very low and
2. s_M could, by chance, be very low.

Intuitively, it makes sense that the probability of being within 1.96 standard errors of the mean should be smaller than in the case when the standard deviation is known (and cannot be underestimated). But exactly how much smaller? Fortunately, the way to work out this type of problem was solved in the early 20th century by W. S. Gosset who determined the distribution of a mean divided by an estimate of its standard error. This distribution is called the Student's t distribution or sometimes just the t distribution. Gosset worked out the t distribution and associated statistical tests while working for a brewery in Ireland. Because of a contractual agreement with the brewery, he published the article under the pseudonym "Student." That is why the t test is called the "Student's t test."

The t distribution is very similar to the normal distribution when the estimate of variance is based on many degrees of freedom, but has relatively more scores in its tails when there are fewer degrees of freedom. Figure $7.8.1$ shows t distributions with 2, 4, and 10 degrees of freedom and the standard normal distribution. Notice that the normal distribution has relatively more scores in the center of the distribution and the t distribution has relatively more in the tails. The t distribution is therefore leptokurtic. The t distribution approaches the normal distribution as the degrees of freedom increase.

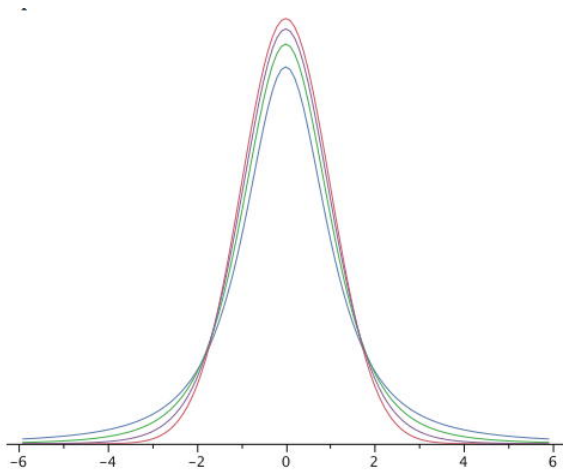


Figure $7.8.1$: A comparison of t distributions with 2, 4, and 10 df and the standard normal distribution. The distribution with the lowest peak is the 2 df distribution, the next lowest is 4 df, the lowest after that is 10 df, and the highest is the standard normal distribution

Since the t distribution is leptokurtic, the percentage of the distribution within 1.96 standard deviations of the mean is less than the 95% for the normal distribution. Table $7.8.1$ shows the number of standard deviations from the mean required to contain 95% and

99% of the area of the t distribution for various degrees of freedom. These are the values of t that you use in a confidence interval. The corresponding values for the normal distribution are 1.96 and 2.58 respectively. Notice that with few degrees of freedom, the values of t are much higher than the corresponding values for a normal distribution and that the difference decreases as the degrees of freedom increase. The values in Table $7.8.1$ can be obtained from the "Find t for a confidence interval" calculator.

Table 7.8.1: *Abbreviated t table*

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Returning to the problem posed at the beginning of this section, suppose you sampled 9 values from a normal population and estimated the standard error of the mean (σ_M) with s_M. What is the probability that M would be within $1.96 s_M$ of μ? Since the sample size is 9, there are $N-1=8df$. From Table $7.8.1$ you can see that with $8df$ the probability is 0.95 that the mean will be within $2.306 s_M$ of μ. The probability that it will be within $1.96 s_M$ of μ is therefore lower than 0.95.

As shown in Figure $7.8.2$, the "t distribution" calculator can be used to find that 0.086 of the area of a t distribution is more than 1.96 standard deviations from the mean, so the probability that M would be less than $1.96 s_M$ from μ is $1-0.086=0.914$.

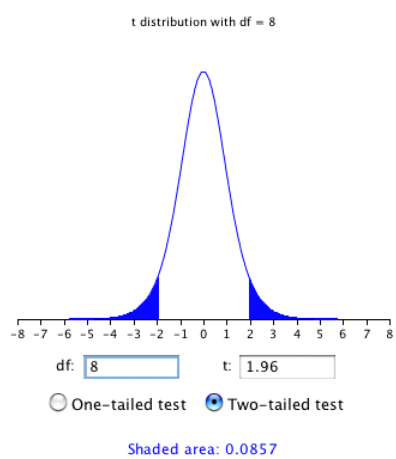


Figure $7.8.2$: Area more than 1.96 standard deviations from the mean in a t distribution with $8df$. Note that the two-tailed button is selected so that the area in both tails will be included

As expected, this probability is less than 0.95 that would have been obtained if σ_M had been known instead of estimated.

7.9: Confidence Interval Simulation

Learning Objectives

- Develop a basic understanding of the properties of a sampling distribution based on the properties of the population

Instructions

This simulation illustrates confidence intervals. For each run of the simulation, 100 sample experiments are conducted and a confidence interval on the mean is computed for each experiment. In each experiment, scores are sampled from a population with a mean of 50 and a standard deviation of 10. Therefore the parameter being estimated is always 50. Both 95% and 99% confidence intervals are computed for each experiment. If the 95% confidence interval contains the population mean of 50, then the confidence interval is shown as an orange line. If the interval does not contain 50, it is shown as a red line. The 99% confidence intervals are shown by extending the 95% intervals. The extension is in blue if the 99% interval contains 50 and in white if it does not. You can choose to make the sample size for each experiment 10, 15, or 20. One hundred simulated experiments are conducted when you click the "Sample" button. The cumulative results are shown at the bottom of the display. You can reset the cumulative results by clicking the "Clear" button.

Illustrated Instructions

The demonstration generates confidence intervals for sample experiments taken from a population with a mean of 50 and a standard deviation of 10. You can choose from various sample sizes but as can be seen from the figure below the default size is 10. Each time you click the sample button confidence intervals for 100 experiments are generated and displayed on the graph on the left.

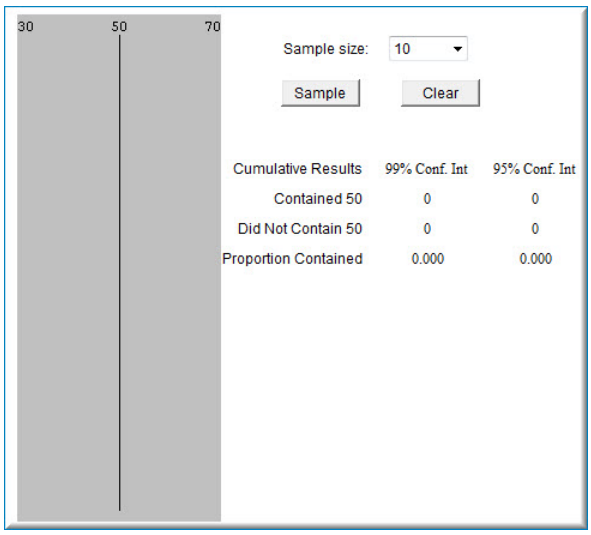


Figure 7.9.1: Confidence interval simulation

The figure below displays the results of 300 experiments with a sample size of 10. The 95% confidence intervals that contain the mean of 50 are shown in orange and the those that do not are shown in red. The 99% confidence intervals are shown in blue if they contain 50 and white if they do not.

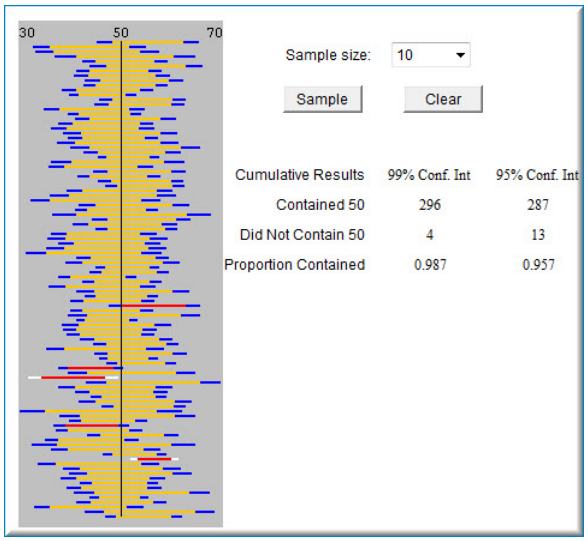


Figure 7.9.2: Confidence interval simulation

7.10: Difference between Means

Learning Objectives

- State the assumptions for computing a confidence interval on the difference between means
- Compute a confidence interval on the difference between means
- Format data for computer analysis

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. We take as an example the data from the "Animal Research" case study. In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 7.10.1.

Table 7.10.1: Means and Variances in Animal Research study

Condition	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However, the gender difference in this particular sample is not very important. What is important is the difference in the population. The difference in sample means is used to estimate the difference in population means. The accuracy of the estimate is revealed by a confidence interval.

In order to construct a confidence interval, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value.

The consequences of violating these assumptions are discussed in a later section. For now, suffice it to say that small-to-moderate violations of assumptions 1 and 2 do not make much difference.

A confidence interval on the difference between means is computed using the following formula:

$$\text{Lower Limit} = M_1 - M_2 - (t_{CL})(S_{M_1-M_2}) \tag{7.10.1}$$

$$\text{Upper Limit} = M_1 - M_2 + (t_{CL})(S_{M_1-M_2}) \tag{7.10.2}$$

where $M_1 - M_2$ is the difference between sample means, t_{CL} is the t for the desired level of confidence, and $S_{M_1-M_2}$ is the estimated standard error of the difference between sample means. The meanings of these terms will be made clearer as the calculations are demonstrated.

We continue to use the data from the "Animal Research" case study and will compute a confidence interval on the difference between the mean score of the females and the mean score of the males. For this calculation, we will assume that the variances in each of the two populations are equal.

The first step is to compute the estimate of the standard error of the difference between means ($S_{M_1-M_2}$). Recall from the relevant section in the chapter on sampling distributions that the formula for the standard error of the difference in means in the population is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}} \tag{7.10.3}$$

In order to estimate this quantity, we estimate σ^2 and use that estimate in place of σ^2. Since we are assuming the population variances are the same, we estimate this variance by averaging our two sample variances. Thus, our estimate of variance is computed using the following formula:

$$MSE = \frac{s_1^2 + s_2^2}{2} \tag{7.10.4}$$

where MSE is our estimate of σ^2. In this example,

$$MSE = \frac{2.743 + 2.985}{2} = 2.864 \tag{7.10.5}$$

Note that MSE stands for "**mean square error**" and is the mean squared deviation of each score from its group's mean.

Since n (the number of scores in each condition) is 17,

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805 \tag{7.10.6}$$

The next step is to find the t to use for the confidence interval (t_{CL}). To calculate t_{CL}, we need to know the degrees of freedom. The degrees of freedom is the number of independent estimates of variance on which MSE is based. This is equal to $(n_1 - 1) + (n_2 - 1)$ where n_1 is the sample size of the first group and n_2 is the sample size of the second group. For this example, $n_1 = n_2 = 17$. When $n_1 = n_2$, it is conventional to use "n" to refer to the sample size of each group. Therefore, the degrees of freedom is $16 + 16 = 32$.

Calculator: Find t for confidence interval

From either the above calculator or a t table, you can find that the t for a 95% confidence interval for $32df$ is 2.037.

We now have all the components needed to compute the confidence interval. First, we know the difference between means:

$$M_1 - M_2 = 5.353 - 3.882 = 1.471 \tag{7.10.7}$$

We know the standard error of the difference between means is

$$S_{M_1 - M2} = 0.5805 \tag{7.10.8}$$

and that the t for the 95% confidence interval with $32df$ is

$$t_{CL} = 2.037 \tag{7.10.9}$$

Therefore, the 95% confidence interval is

$$\text{Lower Limit} = 1.471 - (2.037)(0.5805) = 0.29 \tag{7.10.10}$$

$$\text{Upper Limit} = 1.471 + (2.037)(0.5805) = 2.65 \tag{7.10.11}$$

We can write the confidence interval as:

$$0.29 \le \mu_f - \mu_m \le 2.65 \tag{7.10.12}$$

where μ_f is the population mean for females and μ_m is the population mean for males. This analysis provides evidence that the mean for females is higher than the mean for males, and that the difference between means in the population is likely to be between 0.29 and 2.65.

Formatting data for Computer Analysis

Most computer programs that compute t tests require your data to be in a specific form. Consider the data in Table $7.10.2$

Table $7.10.2$: Example Data

Group 1	Group 2
3	5
4	6
5	7

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. For the data in Table 7.10.2, the reformatted data look as follows:

Table 7.10.3: Reformatted Data

G	Y
1	3
1	4
1	5
2	5
2	6
2	7

To use Analysis Lab to do the calculations, you would copy the data and then

Click the "Enter/Edit User Data" button. (You may be warned that for security reasons you must use the keyboard shortcut for pasting data.)

1. Paste your data.
2. Click "Accept Data."
3. Set the Dependent Variable to Y.
4. Set the Grouping Variable to G.
5. Click the t-test confidence interval button.

The 95% confidence interval on the difference between means extends from -4.267 to 0.267.

Computations for Unequal Sample Sizes (optional)

The calculations are somewhat more complicated when the sample sizes are not equal. One consideration is that MSE, the estimate of variance, counts the sample with the larger sample size more than the sample with the smaller sample size. Computationally this is done by computing the sum of squares error (SSE) as follows:

$$SSE = \sum(X - M_1)^2 + \sum(X - M_2)^2 \tag{7.10.13}$$

where M_1 is the mean for group 1 and M_2 is the mean for group 2. Consider the following small example:

Table 7.10.4: Example Data

Group 1	Group 2
3	2
4	4
5	

$$M_1 = 4 \;\; and \;\; M_2 = 3 \tag{7.10.14}$$

$$SSE = (3-4)^2 + (4-4)^2 + (5-4)^2 + (2-3)^2 + (4-3)^2 = 4 \tag{7.10.15}$$

Then, MSE is computed by:

$$MSE = \frac{SSE}{df} \tag{7.10.16}$$

where the degrees of freedom (df) is computed as before:

$$df = (n_1 - 1) + (n_2 - 1) = (3 - 1) + (2 - 1) = 3 \tag{7.10.17}$$

$$MSE = \frac{SSE}{df} = \frac{4}{3} = 1.333 \tag{7.10.18}$$

The formula

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{n}} \tag{7.10.19}$$

is replaced by

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{n_h}} \tag{7.10.20}$$

where n_h is the harmonic mean of the sample sizes and is computed as follows:

$$n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{2}{\frac{1}{3} + \frac{1}{2}} = 2.4 \tag{7.10.21}$$

and

$$S_{M_1 - M_2} = \sqrt{\frac{(2)(1.333)}{2.4}} = 1.054 \tag{7.10.22}$$

t_{CL} for $3df$ and the 0.05 level equals 3.182.

Therefore the 95% confidence interval is

$$\text{Lower Limit} = 1 - (3.182)(1.054) = -2.35 \tag{7.10.23}$$

$$\text{Upper Limit} = 1 + (3.182)(1.054) = 4.35 \tag{7.10.24}$$

We can write the confidence interval as:

$$-2.35 \leq \mu_1 - \mu_2 \leq 4.35 \tag{7.10.25}$$

7.11: Correlation

Learning Objectives

- State the standard error of z'
- Compute a confidence interval on ρ

The computation of a confidence interval on the population value of Pearson's correlation (ρ) is complicated by the fact that the sampling distribution of r is not normally distributed. The solution lies with Fisher's z' transformation described in the section on the sampling distribution of Pearson's r. The steps in computing a confidence interval for ρ are:

1. Convert r to z'.
2. Compute a confidence interval in terms of z'.
3. Convert the confidence interval back to r.

Let's take the data from the case study Animal Research as an example. In this study, students were asked to rate the degree to which they thought animal research is wrong and the degree to which they thought it is necessary. As you might have expected, there was a negative relationship between these two variables: the more that students thought animal research is wrong, the less they thought it is necessary. The correlation based on 34 observations is -0.654. The problem is to compute a 95% confidence interval on ρ based on this r of -0.654.

The conversion of r to z' can be done using a calculator. This calculator shows that the z' associated with an r of -0.654 is -0.78.

The sampling distribution of z' is approximately normally distributed and has a standard error of

$$\frac{1}{\sqrt{N-3}} \tag{7.11.1}$$

For this example, $N = 34$ and therefore the standard error is 0.180. The Z for a 95% confidence interval ($Z_{0.95}$) is 1.96, as can be found using the normal distribution calculator (setting the shaded area to 0.95 and clicking on the "Between" button). The confidence interval is therefore computed as:

$$\text{Lower limit} = -0.775 - (1.96)(0.18) = -1.13 \tag{7.11.2}$$

$$\text{Upper limit} = -0.775 + (1.96)(0.18) = -0.43 \tag{7.11.3}$$

The final step is to convert the endpoints of the interval back to r using a calculator. The r associated with a z' of -1.13 is -0.81 and the r associated with a z' of -0.43 is -0.40. Therefore, the population correlation (ρ) is likely to be between -0.81 and -0.40. The 95% confidence interval is:

$$-0.81 \le \rho \le -0.40 \tag{7.11.4}$$

To calculate the 99% confidence interval, you use the Z for a 99% confidence interval of 2.58 as follows:

$$\text{Lower limit} = -0.775 - (2.58)(0.18) = -1.24 \tag{7.11.5}$$

$$\text{Upper limit} = -0.775 + (2.58)(0.18) = -0.32 \tag{7.11.6}$$

Converting back to r, the confidence interval is:

$$-0.84 \le \rho \le -0.31 \tag{7.11.7}$$

Naturally, the 99% confidence interval is wider than the 95% confidence interval.

7.12: Proportion

Learning Objectives

- Estimate the population proportion from sample proportions
- Apply the correction for continuity
- Compute a confidence interval

A candidate in a two-person election commissions a poll to determine who is ahead. The pollster randomly chooses 500 registered voters and determines that 260 out of the 500 favor the candidate. In other words, 0.52 of the sample favors the candidate. Although this point estimate of the proportion is informative, it is important to also compute a confidence interval. The confidence interval is computed based on the mean and standard deviation of the sampling distribution of a proportion. The formulas for these two parameters are shown below:

$$\mu_p = \pi \tag{7.12.1}$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{N}} \tag{7.12.2}$$

Since we do not know the population parameter π, we use the sample proportion p as an estimate. The estimated standard error of p is therefore

$$s_p = \sqrt{\frac{p(1-p)}{N}} \tag{7.12.3}$$

We start by taking our statistic (p) and creating an interval that ranges $(Z_{0.95})(s_p)$ in both directions, where $Z_{0.95}$ is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area (see the section on the confidence interval for the mean). The value of $Z_{0.95}$ is computed with the normal calculator and is equal to 1.96. We then make a slight adjustment to correct for the fact that the distribution is discrete rather than continuous.

Normal Distribution Calculator

s_p is calculated as shown below:

$$s_p = \sqrt{\frac{(0.52)(1-0.52)}{300}} = 0.0223 \tag{7.12.4}$$

To correct for the fact that we are approximating a discrete distribution with a continuous distribution (the normal distribution), we subtract $0.5/N$ from the lower limit and add $0.5/N$ to the upper limit of the interval. Therefore the confidence interval is

$$p \pm Z_{0.95}\sqrt{\frac{p(1-p)}{N}} \pm \frac{0.5}{N} \tag{7.12.5}$$

$$\text{Lower limit}: 0.52 - (1.96)(0.0223) - 0.001 = 0.475 \tag{7.12.6}$$

$$\text{Upper limit}: 0.52 + (1.96)(0.0223) + 0.001 = 0.565 \tag{7.12.7}$$

$$0.475 \le \pi \le 0.565 \tag{7.12.8}$$

Since the interval extends 0.045 in both directions, the margin of error is 0.045. In terms of percent, between 47.5% and 56.5% of the voters favor the candidate and the margin of error is 4.5%. Keep in mind that the margin of error of 4.5% is the margin of error for the percent favoring the candidate and not the margin of error for the difference between the percent favoring the candidate and the percent favoring the opponent. The margin of error for the difference is 9%, twice the margin of error for the individual percent. Keep this in mind when you hear reports in the media; the media often get this wrong.

7.13: Statistical Literacy

Learning Objectives

- No "Large Conclusions" from "Tiny" Samples?

In July of 2011, Gene Munster of Piper Jaffray reported the results of a survey in a note to clients. This research was reported throughout the media. Perhaps the fullest description was presented on the CNNMoney website (A service of CNN, Fortune, and Money) in an article entitled “Survey: iPhone retention 94% vs. Android 47%.” The data were collected by asking people in food courts and baseball stadiums what their current phone was and what phone they planned to buy next. The data were collected in the summer of 2011. Below is a portion of the data:

Table 7.13.1: Sample phone retention data

Phone	Keep	Change	Proportion
iPhone	58	4	0.94
Android	17	19	0.47

The article contains the strong caution: “It's only a tiny sample, so large conclusions must not be drawn.” This caution appears to be a welcome change from the overstating of findings typically found in the media. But has this report understated the importance of the study? Perhaps it is valid to draw some "large conclusions."

Example 7.13.1: what do you think?

Is it possible to conclude the vast majority of iPhone owners in the population sampled plan to buy another iPhone or is the sample size too small to justify this conclusion?

Solution

The confidence interval on the proportion extends from 0.87 to 1.0 (some methods give the interval from 0.85 to 0.97). Even the lower bound indicates the vast majority of iPhone owners plan to buy another iPhone. A strong conclusion can be made even with this sample size.

7.E: Estimation (Exercises)

You may want to use the Analysis Lab and various calculators for some of these exercises.

Calculators:

- Inverse t Distribution: Finds t for a confidence interval.
- t Distribution: Computes areas of the t distribution.
- Fisher's r to z': Computes transformations in both directions.
- Inverse Normal Distribution: Use for confidence intervals.

General Questions

Q1

When would the mean grade in a class on a final exam be considered a statistic? When would it be considered a parameter? (relevant section)

Q2

Define bias in terms of expected value. (relevant section)

Q3

Is it possible for a statistic to be unbiased yet very imprecise? How about being very accurate but biased? (relevant section)

Q4

Why is a 99% confidence interval wider than a 95% confidence interval? (relevant section & relevant section)

Q5

When you construct a 95% confidence interval, what are you 95% confident about? (relevant section)

Q6

What is the difference in the computation of a confidence interval between cases in which you know the population standard deviation and cases in which you have to estimate it? (relevant section & relevant section)

Q7

Assume a researcher found that the correlation between a test he or she developed and job performance was 0.55 in a study of 28 employees. If correlations under 0.35 are considered unacceptable, would you have any reservations about using this test to screen job applicants? (relevant section)

Q8

What is the effect of sample size on the width of a confidence interval? (relevant section & relevant section)

Q9

How does the t distribution compare with the normal distribution? How does this difference affect the size of confidence intervals constructed using z relative to those constructed using t? Does sample size make a difference? (relevant section)

Q10

The effectiveness of a blood-pressure drug is being investigated. How might an experimenter demonstrate that, on average, the reduction in systolic blood pressure is 20 or more? (relevant section & relevant section)

Q11

A population is known to be normally distributed with a standard deviation of 2.8.

a. Compute the 95% confidence interval on the mean based on the following sample of nine: $8, 9, 10, 13, 14, 16, 17, 20, 21$
b. Now compute the 99% confidence interval using the same data. (relevant section)

Q12

A person claims to be able to predict the outcome of flipping a coin. This person is correct $16/25$ times. Compute the 95% confidence interval on the proportion of times this person can predict coin flips correctly. What conclusion can you draw about this test of his ability to predict the future? (relevant section)

Q13

What does it mean that the variance (computed by dividing by N) is a biased statistic? (relevant section)

Q14

A confidence interval for the population mean computed from an N of 16 ranges from 12 to 28. A new sample of 36 observations is going to be taken. You can't know in advance exactly what the confidence interval will be because it depends on the random sample. Even so, you should have some idea of what it will be. Give your best estimation. (relevant section)

Q15

You take a sample of 22 from a population of test scores, and the mean of your sample is 60.

a. You know the standard deviation of the population is 10. What is the 99% confidence interval on the population mean.
b. Now assume that you do not know the population standard deviation, but the standard deviation in your sample is 10. What is the 99% confidence interval on the mean now? (relevant section)

Q16

You read about a survey in a newspaper and find that 70% of the 250 people sampled prefer Candidate A. You are surprised by this survey because you thought that more like 50% of the population preferred this candidate. Based on this sample, is 50% a possible population proportion? Compute the 95% confidence interval to be sure. (relevant section)

Q17

Heights for teenage boys and girls were calculated. The mean height for the sample of 12 boys was 174 cm and the variance was 62. For the sample of 12 girls, the mean was 166 cm and the variance was 65.

a. What is the 95% confidence interval on the difference between population means?
b. What is the 99% confidence interval on the difference between population means?
c. Do you think the mean difference in the population could be about 5? Why or why not? (relevant section)

Q18

You were interested in how long the average psychology major at your college studies per night, so you asked 10 psychology majors to tell you the amount they study. They told you the following times: $2, 1.5, 3, 2, 3.5, 1, 0.5, 3, 2, 4$

a. Find the 95% confidence interval on the population mean.
b. Find the 90% confidence interval on the population mean. (relevant section)

Q19

True/false: As the sample size gets larger, the probability that the confidence interval will contain the population mean gets higher. (relevant section & relevant section)

Q20

True/false: You have a sample of 9 men and a sample of 8 women. The degrees of freedom for the t value in your confidence interval on the difference between means is 16. (relevant section & relevant section)

Q21

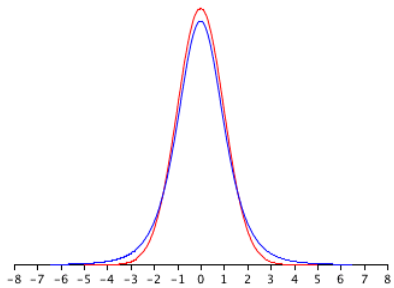
True/false: Greek letters are used for statistics as opposed to parameters. (relevant section)

Q22

True/false: In order to construct a confidence interval on the difference between means, you need to assume that the populations have the same variance and are both normally distributed. (relevant section)

Q23

True/false: The red distribution represents the t distribution and the blue distribution represents the normal distribution. (relevant section)



Questions from Case Studies

The following questions are from the Angry Moods (AM) case study.

Q24

(AM#6c) Is there a difference in how much males and females use aggressive behavior to improve an angry mood? For the "Anger-Out" scores, compute a 99% confidence interval on the difference between gender means. (relevant section)

Q25

(AM#10) Calculate the 95% confidence interval for the difference between the mean Anger-In score for the athletes and non-athletes. What can you conclude? (relevant section)

Q26

Find the 95% confidence interval on the population correlation between the Anger-Out and Control-Out scores. (relevant section)

The following questions are from the Flatulence (F) case study.

Q27

(F#8) Compare men and women on the variable "perday." Compute the 95% confidence interval on the difference between means. (relevant section)

Q28

(F#10) What is the 95% confidence interval of the mean time people wait before farting in front of a romantic partner. (relevant section)

The following questions use data from the Animal Research (AR) case study.

Q29

(AR#3) What percentage of the women studied in this sample strongly agreed (gave a rating of 7) that using animals for research is wrong?

Q30

Use the proportion you computed in #29. Compute the 95% confidence interval on the population proportion of women who strongly agree that animal research is wrong. (relevant section)

Q31

Compute a 95% confidence interval on the difference between the gender means with respect to their beliefs that animal research is wrong. (relevant section)

The following question is from the ADHD Treatment (AT) case study.

Q32

(AT#8) What is the correlation between the participants' correct number of responses after taking the placebo and their correct number of responses after taking 0.60 mg/kg of MPH? Compute the 95% confidence interval on the population correlation. (relevant section)

The following question is from the Weapons and Aggression (WA) case study.

Q33

(WA#4) Recall that the hypothesis is that a person can name an aggressive word more quickly if it is preceded by a weapon word prime than if it is preceded by a neutral word prime. The first step in testing this hypothesis is to compute the difference between

i. the naming time of aggressive words when preceded by a neutral word prime and
ii. the naming time of aggressive words when preceded by a weapon word prime separately for each of the 32 participants. That is, compute an $-aw$ for each participant.

a. Would the hypothesis of this study be supported if the difference were positive or if it were negative?
b. What is the mean of this difference score? (relevant section)
c. What is the standard deviation of this difference score? (relevant section)
d. What is the 95% confidence interval of the mean difference score? (relevant section)
e. What does the confidence interval computed in (d) say about the hypothesis.

The following question is from the Diet and Health (WA) case study.

Q34

Compute a 95% confidence interval on the proportion of people who are healthy on the AHA diet.

	Cancers	Deaths	Nonfatal illness	Healthy	Total
AHA					

15 24 25 239 303 Mediterranean 7 14 8 273 302 Total 22 38 33 512 605

The following questions are from (reproduced with permission) Visit the site

Q35

Suppose that you take a random sample of $10,000$ Americans and find that $1,111$ are left-handed. You perform a test of significance to assess whether the sample data provide evidence that more than 10% of all Americans are left-handed, and you calculate a test statistic of 3.70 and a p-value of 0.0001. Furthermore, you calculate a 99% confidence interval for the proportion of left-handers in America to be $(0.103, 0.119)$ Consider the following statements: The sample provides strong evidence that more than 10% of all Americans are left-handed. The sample provides evidence that the proportion of left-handers in America is much larger than 10%. Which of these two statements is the more appropriate conclusion to draw? Explain your answer based on the results of the significance test and confidence interval.

Q36

A student wanted to study the ages of couples applying for marriage licenses in his county. He studied a sample of 94 marriage licenses and found that in 67 cases the husband was older than the wife. Do the sample data provide strong evidence that the husband is usually older than the wife among couples applying for marriage licenses in that county? Explain briefly and justify your answer.

Q37

Imagine that there are 100 different researchers each studying the sleeping habits of college freshmen. Each researcher takes a random sample of size 50 from the same population of freshmen. Each researcher is trying to estimate the mean hours of sleep that

freshmen get at night, and each one constructs a 95% confidence interval for the mean. Approximately how many of these 100 confidence intervals will NOT capture the true mean?

1. None
2. 1 or 2
3. 3 to 7
4. about half
5. 95 to 100
6. other

Selected Answers

S11

a. $(12.39, 16.05)$

S12

$(0.43, 0.85)$

S15

b. $(53.96, 66.04)$

S17

a. $(1.25, 14.75)$

S18

a. $(1.45, 3.05)$

S26

$(-0.713, -0.414)$

S27

$(-0.98, 3.09)$

S29

41%

S33

b. 7.16

CHAPTER OVERVIEW

8: Tests of Means

Many, if not most experiments are designed to compare means. The experiment may involve only one sample mean that is to be compared to a specific value. Or the experiment could be testing differences among many different experimental conditions, and the experimenter could be interested in comparing each mean with each of the other means. This chapter covers methods of comparing means in many different experimental situations. The topics covered here in sections E, F, I, and J are typically covered in other texts in a chapter on Analysis of Variance. We prefer to cover them here since they bear no necessary relationship to analysis of variance. As discussed by Wilkinson (1999), it is not logical to consider the procedures in this chapter as tests to be performed subsequent to an analysis of variance. Nor is it logical to call them post-hoc tests as some computer programs do.

Thumbnail: Student's t-distribution with 2 degrees of freedom. (CC BY-SA 3.0; IkamusumeFan).

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

8.1: Testing a Single Mean

Learning Objectives

- Compute the probability of a sample mean being at least as high as a specified value when σ is known
- Compute a two-tailed probability
- Compute the probability of a sample mean being at least as high as a specified value when σ is estimated
- State the assumptions required for item 3 above

This section shows how to test the null hypothesis that the population mean is equal to some hypothesized value. For example, suppose an experimenter wanted to know if people are influenced by a subliminal message and performed the following experiment. Each of nine subjects is presented with a series of 100 pairs of pictures. As a pair of pictures is presented, a subliminal message is presented suggesting the picture that the subject should choose. The question is whether the (population) mean number of times the suggested picture is chosen is equal to 50. In other words, the null hypothesis is that the population mean (μ) is 50. The (hypothetical) data are shown in Table $8.1.1$. The data in Table $8.1.1$ have a sample mean (M) of 51. Thus the sample mean differs from the hypothesized population mean by 1.

Table $8.1.1$: Distribution of scores

Frequency
45
48
49
49
51
52
53
55
57

The significance test consists of computing the probability of a sample mean differing from μ by one (the difference between the hypothesized population mean and the sample mean) or more. The first step is to determine the sampling distribution of the mean. As shown in a previous section, the mean and standard deviation of the sampling distribution of the mean are

$$\mu_M = \mu \tag{8.1.1}$$

and

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{8.1.2}$$

respectively. It is clear that $\mu_M = 50$. In order to compute the standard deviation of the sampling distribution of the mean, we have to know the population standard deviation (σ).

The current example was constructed to be one of the few instances in which the standard deviation is known. In practice, it is very unlikely that you would know σ and therefore you would use s, the sample estimate of σ. However, it is instructive to see how the probability is computed if σ is known before proceeding to see how it is calculated when σ is estimated.

For the current example, if the null hypothesis is true, then based on the binomial distribution, one can compute that variance of the number correct is

$$\sigma^2 = N\pi(1-\pi) = 100(0.5)(1-0.5) = 25 \tag{8.1.3}$$

Therefore, $\sigma = 5$. For a σ of 5 and an N of 9, the standard deviation of the sampling distribution of the mean is $5/3 = 1.667$. Recall that the standard deviation of a sampling distribution is called the standard error.

To recap, we wish to know the probability of obtaining a sample mean of 51 or more when the sampling distribution of the mean has a mean of 50 and a standard deviation of 1.667. To compute this probability, we will make the assumption that the sampling distribution of the mean is normally distributed. We can then use the normal distribution calculator as shown in Figure $8.1.1$.

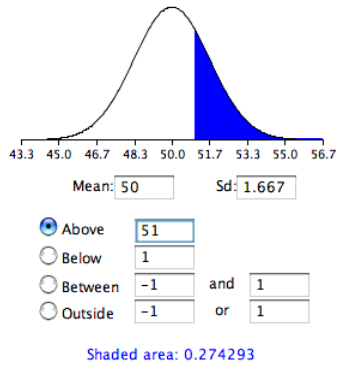


Figure $8.1.1$: Probability of a sample mean being 51 or greater

Notice that the mean is set to 50, the standard deviation to 1.667, and the area above 51 is requested and shown to be 0.274.

Therefore, the probability of obtaining a sample mean of 51 or larger is 0.274. Since a mean of 51 or higher is not unlikely under the assumption that the subliminal message has no effect, the effect is not significant and the null hypothesis is not rejected.

The test conducted above was a one-tailed test because it computed the probability of a sample mean being one or more points higher than the hypothesized mean of 50 and the area computed was the area above 51. To test the two-tailed hypothesis, you would compute the probability of a sample mean differing by one or more in either direction from the hypothesized mean of 50. You would do so by computing the probability of a mean being less than or equal to 49 or greater than or equal to 51.

The results of the normal distribution calculator are shown in Figure $8.1.2$.

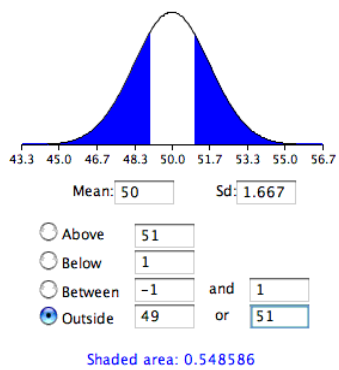


Figure $8.1.2$: Probability of a sample mean being less than or equal to 49 or greater than or equal to 51

As you can see, the probability is 0.548 which, as expected, is twice the probability of 0.274 shown in Figure $8.1.1$.

Before normal calculators such as the one illustrated above were widely available, probability calculations were made based on the standard normal distribution. This was done by computing Z based on the formula

$$Z = \frac{M - \mu}{\sigma_M} \tag{8.1.4}$$

where Z is the value on the standard normal distribution, M is the sample mean, μ is the hypothesized value of the mean, and σ_M is the standard error of the mean. For this example, $Z = (51 - 50)/1.667 = 0.60$. Use the normal calculator, with a mean of 0 and a standard deviation of 1, as shown below.

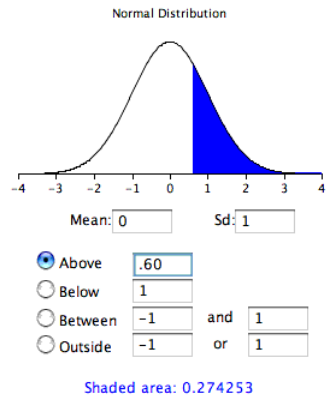


Figure 8.1.3: *Calculation using the standardized normal distribution*

Notice that the probability (the shaded area) is the same as previously calculated (for the one-tailed test).

As noted, in real-world data analysis it is very rare that you would know σ and wish to estimate μ. Typically σ is not known and is estimated in a sample by s, and σ_M is estimated by s_M.

Example 8.1.1: adhd treatment

Consider the data in the "ADHD Treatment" case study. These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. Table $8.1.2$ shows the data for the placebo (0 mg) and highest dosage level (0.6 mg) of methylphenidate. Of particular interest here is the column labeled "Diff" that shows the difference in performance between the 0.6 mg ($D60$) and the 0 mg ($D0$) conditions. These difference scores are positive for children who performed better in the 0.6 mg condition than in the control condition and negative for those who scored better in the control condition. If methylphenidate has a positive effect, then the mean difference score in the population will be positive. The null hypothesis is that the mean difference score in the population is 0.

***Table* 8.1.2:** *DOG scores as a function of dosage*

D0	D60	Diff
57	62	5
27	49	22
32	30	-2
31	34	3
34	38	4
38	36	-2
71	77	6
33	51	18
34	45	11

53	42	-11
36	43	7
42	57	15
26	36	10
52	58	6
36	35	-1
55	60	5
36	33	-3
42	49	7
36	33	-3
54	59	5
34	35	1
29	37	8
33	45	12
33	29	-4

Solution

To test this null hypothesis, we compute t using a special case of the following formula:

$$t = \frac{\text{statistic-hypothesized value}}{\text{estimated standard error of the statistic}} \tag{8.1.5}$$

The special case of this formula applicable to testing a single mean is

$$t = \frac{M - \mu}{s_M} \tag{8.1.6}$$

where t is the value we compute for the significance test, M is the sample mean, μ is the hypothesized value of the population mean, and s_M is the estimated standard error of the mean. Notice the similarity of this formula to the formula for Z.

In the previous example, we assumed that the scores were normally distributed. In this case, it is the population of difference scores that we assume to be normally distributed.

The mean (M) of the $N = 24$ difference scores is 4.958, the hypothesized value of μ is 0, and the standard deviation (s) is 7.538. The estimate of the standard error of the mean is computed as:

$$s_M = \frac{s}{\sqrt{N}} = \frac{7.5382}{\sqrt{24}} = 1.54 \tag{8.1.7}$$

Therefore,

$$t = \frac{4.96}{1.54} = 3.22 \tag{8.1.8}$$

The probability value for t depends on the degrees of freedom. The number of degrees of freedom is equal to $N - 1 = 23$. As shown below, the t distribution calculator finds that the probability of a t less than -3.22 or greater than 3.22 is only 0.0038. Therefore, if the drug had no effect, the probability of finding a difference between means as large or larger (in either direction) than the difference found is very low. Therefore the null hypothesis that the population mean difference score is zero can be rejected. The conclusion is that the population mean for the drug condition is higher than the population mean for the placebo condition.

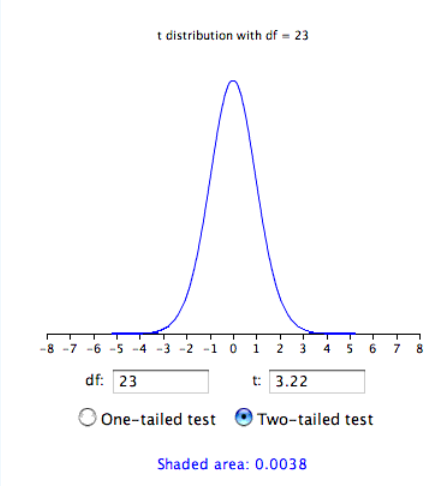


Figure 8.1.4: t distribution

Review of Assumptions

1. Each value is sampled independently from each other value.
2. The values are sampled from a normal distribution.

8.2: t Distribution Demo

Learning Objectives

- State how the degrees of freedom affect the difference between the t and normal distributions

Instructions

This demonstration allows you to compare the t distribution to the standard normal distribution. At the start, the standard normal distribution is compared to a t distribution with three degrees of freedom. You can change the degrees of freedom of the t distribution with the slider. The 3 and the 50 mark the ends of the slider. The "current" degrees of freedom are shown at the bottom where it says "t distribution with $df = 3$." As you change the slider, the df are shown by this last line. The "zoom in" button allows you to change the scale of the graph to see the tails of the distribution in more detail.

Illustrated Instructions

Video Demo

In the video the slider is used to increase the degrees of freedom. Notice how the t-distribution change as the degrees of freedom increase. The video concludes with the tails of the tails of the graphs being zoomed into via the "zoom in" button and the degrees of freedom being decreased.

8.3: Difference between Two Means

Learning Objectives

- State the assumptions for testing the difference between two means
- Estimate the population variance assuming homogeneity of variance
- Compute the standard error of the difference between means
- Compute t and p for the difference between means
- Format data for computer analysis

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. This section covers how to test for differences between means from two separate groups of subjects. A later section describes how to test for differences between the means of two conditions in designs where only one group of subjects is used and each subject is tested in each condition.

We take as an example the data from the "Animal Research" case study. In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table $8.3.1$.

Table* $8.3.1$*: *Means and Variances in Animal Research study*

Group	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However, the gender difference in this particular sample is not very important. What is important is whether there is a difference in the population means.

In order to test whether there is a difference between population means, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the scores are not independent. The analysis of data with two scores per subject is shown in the section on the correlated t test later in this chapter.

The consequences of violating the first two assumptions are investigated in the simulation in the next section. For now, suffice it to say that small-to-moderate violations of assumptions 1 and 2 do not make much difference. It is important not to violate assumption 3.

We saw the following general formula for significance testing in the section on testing a single mean:

$$t = \frac{\text{statistic-hypothesized value}}{\text{estimated standard error of the statistic}} \tag{8.3.1}$$

In this case, our statistic is the difference between sample means and our hypothesized value is 0. The hypothesized value is the null hypothesis that the difference between population means is 0.

We continue to use the data from the "Animal Research" case study and will compute a significance test on the difference between the mean score of the females and the mean score of the males. For this calculation, we will make the three assumptions specified above.

The first step is to compute the statistic, which is simply the difference between means.

$$M_1 - M_2 = 5.3529 - 3.8824 = 1.4705 \tag{8.3.2}$$

Since the hypothesized value is 0, we do not need to subtract it from the statistic.

The next step is to compute the estimate of the standard error of the statistic. In this case, the statistic is the difference between means, so the estimated standard error of the statistic is ($S_{M_1-M_2}$). Recall from the relevant section in the chapter on sampling distributions that the formula for the standard error of the difference between means is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}} \tag{8.3.3}$$

In order to estimate this quantity, we estimate σ^2 and use that estimate in place of σ^2. Since we are assuming the two population variances are the same, we estimate this variance by averaging our two sample variances. Thus, our estimate of variance is computed using the following formula:

$$MSE = \frac{s_1^2 + s_2^2}{2} \tag{8.3.4}$$

where MSE is our estimate of σ^2. In this example,

$$MSE = \frac{2.743 + 2.985}{2} = 2.864 \tag{8.3.5}$$

Since n (the number of scores in each group) is 17,

$$S_{M_1-M_2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805 \tag{8.3.6}$$

The next step is to compute t by plugging these values into the formula:

$$t = \frac{1.4705}{0.5805} = 2.533 \tag{8.3.7}$$

Finally, we compute the probability of getting a t as large or larger than 2.533 or as small or smaller than -2.533. To do this, we need to know the degrees of freedom. The degrees of freedom is the number of independent estimates of variance on which MSE is based. This is equal to $(n_1 - 1) + (n_2 - 1)$, where n_1 is the sample size of the first group and n_2 is the sample size of the second group. For this example, $n_1 = n_2 = 17$. When $n_1 = n_2$, it is conventional to use "n" to refer to the sample size of each group. Therefore, the degrees of freedom is $16 + 16 = 32$.

Once we have the degrees of freedom, we can use the t distribution calculator to find the probability. Figure $8.3.1$ shows that the probability value for a two-tailed test is 0.0164. The two-tailed test is used when the null hypothesis can be rejected regardless of the direction of the effect. As shown in Figure $8.3.1$, it is the probability of a $t < -2.533$ or a $t > 2.533$.

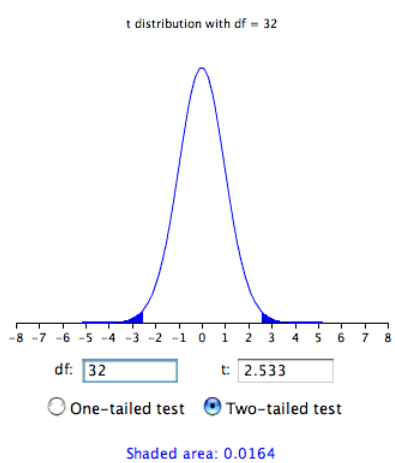


Figure $8.3.1$: The two-tailed probability

The results of a one-tailed test are shown in Figure $8.3.2$. As you can see, the probability value of 0.0082 is half the value for the two-tailed test.

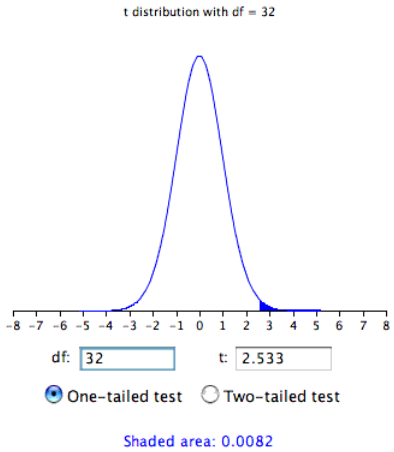


Figure 8.3.2: The one-tailed probability

Formatting Data for Computer Analysis

Most computer programs that compute t tests require your data to be in a specific form. Consider the data in Table 8.3.2.

Table 8.3.2: Example Data

Group 1	Group 2
3	2
4	6
5	8

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table 8.3.2 is shown in Table 8.3.3.

Table 8.3.3: Reformatted Data

G	Y
1	3
1	4
1	5
2	2
2	6
2	8

To use Analysis Lab to do the calculations, you would copy the data and then

1. Click the "Enter/Edit Data" button. (You may be warned that for security reasons you must use the keyboard shortcut for pasting data.)
2. Paste your data.
3. Click "Accept Data."
4. Set the Dependent Variable to Y.
5. Set the Grouping Variable to G.
6. Click the "t-test/confidence interval" button.

The t value is -0.718, the $df = 4$, and $p = 0.512$.

Computations for Unequal Sample Sizes (optional)

The calculations are somewhat more complicated when the sample sizes are not equal. One consideration is that MSE, the estimate of variance, counts the group with the larger sample size more than the group with the smaller sample size. Computationally, this is done by computing the sum of squares error (SSE) as follows:

$$SSE = \sum(X - M_1)^2 + \sum(X - M_2)^2 \tag{8.3.8}$$

where M_1 is the mean for group 1 and M_2 is the mean for group 2. Consider the following small example:

Table 8.3.4: Unequal n

Group 1	Group 2
3	2
4	4
5	

$$M_1 = 4 \text{ and } M_2 = 3 \tag{8.3.9}$$

$$SSE = (3-4)^2 + (4-4)^2 + (5-4)^2 + (2-3)^2 + (4-3)^2 = 4 \tag{8.3.10}$$

Then, MSE is computed by:

$$MSE = \frac{SSE}{df} \tag{8.3.11}$$

where the degrees of freedom (df) is computed as before:

$$df = (n_1 - 1) + (n_2 - 1) = (3-1) + (2-1) = 3 \tag{8.3.12}$$

$$MSE = \frac{SSE}{df} = \frac{4}{3} = 1.333 \tag{8.3.13}$$

The formula

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{n}} \tag{8.3.14}$$

is replaced by

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{\mathfrak{n}_h}} \tag{8.3.15}$$

where $\mathfrak{n}_h$ is the harmonic mean of the sample sizes and is computed as follows:

$$\mathfrak{n}_h = \frac{2}{\frac{1}{\mathfrak{n}_1} + \frac{1}{\mathfrak{n}_2}} = \frac{2}{\frac{1}{3} + \frac{1}{2}} = 2.4 \tag{8.3.16}$$

and

$$S_{M_1 - M_2} = \sqrt{\frac{(2)(1.333)}{2.4}} = 1.054 \tag{8.3.17}$$

Therefore,

$$t = \frac{4-3}{1.054} = 0.949 \tag{8.3.18}$$

and the two-tailed $p = 0.413$.

- **12.3: Difference between Two Means** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

8.4: Robustness Simulation

Learning Objectives

- State when heterogenety of variance can lead to a very high **Type I** error rate
- State the effect of skew of on the **Type I** error rate

Instructions

This demonstration allows you to explore the effects of violating the assumptions of normality and homogeneity of variance. When the simulation starts you see the distributions of two populations. By default, they are both normally distributed, have means of 0 and standard deviations of 1. The default sample size for the simulations is 5 per group. If you push the "simulate" button, $2,000$ simulated experiments are conducted. You can adjust the number of simulations from $2,000$ to $10,000$. A t-test is computed for each experiment and the number of tests that were significant, not significant, and the **Type I** error rate (the proportion significant) are displayed.

Since the null hypothesis is true and all assumptions are met with these default values, the **Type I** error rate should be close to 0.05, especially if you ran a large number of simulations. It will not equal 0.05 because of random variation. However, the larger the number of simulations you run, the closer the **Type I** error rate should come to 0.05.

You can explore the effects of violating the assumptions of the test by making one or both of the distributions skewed and/or by making the standard deviations of the distributions different. You can also explore the effects of sample size and of the significance level used (0.05 or 0.01).

By exploring various distributions, sample sizes, and significance levels, you can get a feeling for how well the test works when its violations are violated. A test that is relatively unaffected by violations of its assumptions is said to be "robust."

Illustrated Instructions

Video Demo

The video below begins by running 2000 simulations with the two populations each with means of 0, standard deviations of 2 no skewness and sample sizes of 5. The video continues by varying different aspect of the distributions and running more simulations. Note the number of significant tests after each set of simulations.

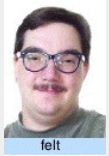
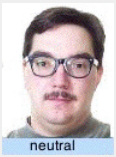
8.5: Pairwise Comparisons

Learning Objectives

- Define pairwise comparison
- Describe the problem with doing t tests among all pairs of means
- Calculate the **Tukey HSD** test
- Explain why the Tukey test should not necessarily be considered a follow-up test

Many experiments are designed to compare more than two conditions. We will take as an example the case study "Smiles and Leniency." In this study, the effect of different types of smiles on the leniency shown to a person was investigated. An obvious way to proceed would be to do a t test of the difference between each group mean and each of the other group means. This procedure would lead to the six comparisons shown in Table $8.5.1$.

Table $8.5.1$: *Six Comparisons among Means*

false vs felt		
false vs miserable		
false vs neutral		
felt vs miserable		
felt vs neutral		
miserable vs neutral		

The problem with this approach is that if you did this analysis, you would have six chances to make a Type I error. Therefore, if you were using the 0.05 significance level, the probability that you would make a **Type I** error on at least one of these comparisons is greater than 0.05. The more means that are compared, the more the **Type I** error rate is inflated. Figure $8.5.1$ shows the number of possible comparisons between pairs of means (pairwise comparisons) as a function of the number of means. If there are only two means, then only one comparison can be made. If there are 12 means, then there are 66 possible comparisons.

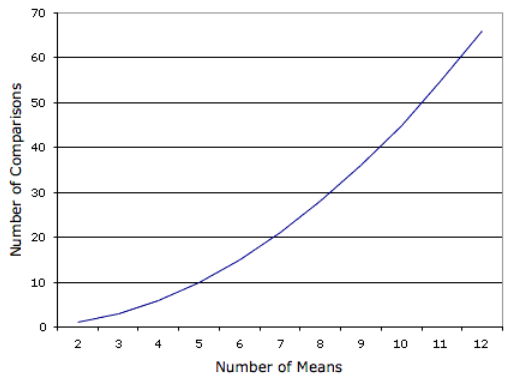


Figure 8.5.1: Number of pairwise comparisons as a function of the number of means

Figure 8.5.2 shows the probability of a **Type I** error as a function of the number of means. As you can see, if you have an experiment with 12 means, the probability is about 0.70 that at least one of the 66 comparisons among means would be significant even if all 12 population means were the same.

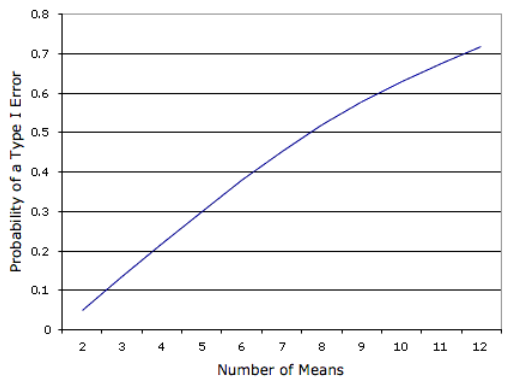


Figure 8.5.2: Probability of a **Type I** error as a function of the number of means

The Tukey Honestly Significant Difference Test

The **Type I** error rate can be controlled using a test called the Tukey Honestly Significant Difference test or **Tukey HSD** for short. The **Tukey HSD** is based on a variation of the t distribution that takes into account the number of means being compared. This distribution is called the **studentized range distribution**.

Let's return to the leniency study to see how to compute the **Tukey HSD** test. You will see that the computations are very similar to those of an independent-groups t test. The steps are outlined below:

1. Compute the means and variances of each group. They are shown below.

Table 8.5.2: Means and variances of each group

Condition	Mean	Variance
False	5.37	3.34
Felt	4.91	2.83
Miserable	4.91	2.11
Neutral	4.12	2.32

2. Compute MSE, which is simply the mean of the variances. It is equal to 2.65.
3. Compute

$$Q = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n}}} \tag{8.5.1}$$

for each pair of means, where M_i is one mean, M_j is the other mean, and n is the number of scores in each group. For these data, there are 34 observations per group. The value in the denominator is 0.279.

4. Compute p for each comparison using the Studentized Range Calculator. The degrees of freedom is equal to the total number of observations minus the number of means. For this experiment, $df = 136 - 4 = 132$.

The tests for these data are shown in Table $8.5.2$.

Table $8.5.2$: *Six Pairwise Comparisons*

Comparison	M_i-M_j	Q	p
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

The only significant comparison is between the false smile and the neutral smile.

It is not unusual to obtain results that on the surface appear paradoxical. For example, these results appear to indicate that

a. the false smile is the same as the miserable smile,
b. the miserable smile is the same as the neutral control, and
c. the false smile is different from the neutral control.

This apparent contradiction is avoided if you are careful not to accept the null hypothesis when you fail to reject it. The finding that the false smile is not significantly different from the miserable smile does not mean that they are really the same. Rather it means that there is not convincing evidence that they are different. Similarly, the non-significant difference between the miserable smile and the control does not mean that they are the same. The proper conclusion is that the false smile is higher than the control and that the miserable smile is either

a. equal to the false smile,
b. equal to the control, or
c. somewhere in-between.

The assumptions of the Tukey test are essentially the same as for an independent-groups t test: normality, homogeneity of variance, and independent observations. The test is quite robust to violations of normality. Violating homogeneity of variance can be more problematical than in the two-sample case since the MSE is based on data from all groups. The assumption of independence of observations is important and should not be violated.

Computer Analysis

For most computer programs, you should format your data the same way you do for an independent-groups t test. The only difference is that if you have, say, four groups, you would code each group as 1, 2, 3, or 4 rather than just 1 or 2.

Although full-featured statistics programs such as **SAS**, **SPSS**, **R**, and others can compute Tukey's test, smaller programs (including Analysis Lab) may not. However, these programs are generally able to compute a procedure known as Analysis of Variance (**ANOVA**). This procedure will be described in detail in a later chapter. Its relevance here is that an ANOVA computes the MSE that is used in the calculation of Tukey's test. For example, the following shows the **ANOVA** summary table for the "Smiles and Leniency" data.

Source	df	SSQ	MS	F	p
Condition	3	27.5349	9.1783	3.4650	0.0182
Error	132	349.6544	2.6489		
Total	135	377.1893			

The column labeled MS stands for "Mean Square" and therefore the value 2.6489 in the "Error" row and the MS column is the "Mean Square Error" or MSE. Recall that this is the same value computed here (2.65) when rounded off.

Tukey's Test Need Not be a Follow-Up to ANOVA

Some textbooks introduce the Tukey test only as a follow-up to an analysis of variance. There is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA (or even know what one is). If you or your instructor do not wish to take our word for this, see the excellent article on this and other issues in statistical analysis by Leland Wilkinson and the APA Board of Scientific Affairs' Task Force on Statistical Inference, published in the *American Psychologist*, August 1999, Vol. *54*, No. 8, 594–604.

Computations for Unequal Sample Sizes (optional)

The calculation of MSE for unequal sample sizes is similar to its calculation in an independent-groups t test. Here are the steps:

1. Compute a Sum of Squares Error (SSE) using the following formula

$$SSE = \sum(X - M_1)^2 + \sum(X - M_2)^2 + \cdots + \sum(X - M_k)^2 \tag{8.5.2}$$

   where M_i is the mean of the ith group and k is the number of groups.
2. Compute the degrees of freedom error (dfe) by subtracting the number of groups (k) from the total number of observations (N). Therefore,

$$dfe = N - k \tag{8.5.3}$$

3. Compute MSE by dividing SSE by dfe:

$$MSE = \frac{SSE}{dfe} \tag{8.5.4}$$

4. For each comparison of means, use the harmonic mean of the $n's$ for the two means ($\mathfrak{n}_\mathfrak{h}$).

All other aspects of the calculations are the same as when you have equal sample sizes.

8.6: Specific Comparisons

Learning Objectives

- Define linear combination
- Specify a linear combination in terms of coefficients
- Do a significance test for a specific comparison

There are many situations in which the comparisons among means are more complicated than simply comparing one mean with another. This section shows how to test these more complex comparisons. The methods in this section assume that the comparison among means was decided on before looking at the data. Therefore these comparisons are called planned comparisons. A different procedure is necessary for unplanned comparisons.

Let's begin with the made-up data from a hypothetical experiment shown in Table $8.6.1$. Twelve subjects were selected from a population of high-self-esteem subjects ($esteem = 1$) and an additional 12 subjects were selected from a population of low-self-esteem subjects ($esteem = 2$). Subjects then performed on a task and (independent of how well they really did) half in each esteem category were told they succeeded ($outcome = 1$) and the other half were told they failed ($outcome = 2$). Therefore, there were six subjects in each of the four esteem/outcome combinations and 24 subjects in all.

After the task, subjects were asked to rate (on a 10-point scale) how much of their outcome (success or failure) they attributed to themselves as opposed to being due to the nature of the task.

Table $8.6.1$: *Data from Hypothetical Experiment*

outcome	esteem	attrib
1	1	7
1	1	8
1	1	7
1	1	8
1	1	9
1	1	5
1	2	6
1	2	5
1	2	7
1	2	4
1	2	5
1	2	6
2	1	4
2	1	6
2	1	5
2	1	4
2	1	7
2	1	3
2	2	9
2	2	8
2	2	9
2	2	8
2	2	7

outcome	esteem	attrib
2	2	6

The means of the four conditions are shown in Table $8.6.2$.

Table $8.6.2$: *Mean ratings of self-attributions of success or failure*

Outcome	Esteem	Mean
Success	High Self-Esteem	7.333
	Low Self-Esteem	5.500
Failure	High Self-Esteem	4.833
	Low Self-Esteem	7.833

There are several questions we can ask about the data. We begin by asking whether, on average, subjects who were told they succeeded differed significantly from subjects who were told they failed. The means for subjects in the success condition are 7.333 for the high-self-esteem subjects and 5.500 for the low-self-esteem subjects. Therefore, the mean for all subjects in the success condition is $\frac{7.3333+5.5000}{2}=6.4167$. Similarly, the mean for all subjects in the failure condition is $\frac{4.8333+7.8333}{2}=6.3333$. The question is: How do we do a significance test for this difference of $6.4167-6.3333=0.083$?

The first step is to express this difference in terms of a linear combination using a set of coefficients and the means. This may sound complex, but it is really pretty easy. We can compute the mean of the success conditions by multiplying each success mean by 0.5 and then adding the result. In other words, we compute

$$(0.5)(7.333)+(0.5)(5.500)=3.67+2.75=6.42 \tag{8.6.1}$$

Similarly, we can compute the mean of the failure conditions by multiplying each "failure" mean by 0.5 and then adding the result:

$$(0.5)(4.833)+(0.5)(7.833)=2.417+3.917=6.33 \tag{8.6.2}$$

The difference between the two means can be expressed as

$$0.5\times 7.333+0.5\times 5.500-(0.5\times 4.833+0.5\times 7.833)=0.5\times 7.333+0.5\times 5.500-0.5\times 4.833-0.5\times 7.833 \tag{8.6.3}$$

We therefore can compute the difference between the "success" mean and the "failure" mean by multiplying each "success" mean by 0.5, each failure mean by -0.5, and adding the results. In Table $8.6.3$, the coefficient column is the multiplier and the product column is the result of the multiplication. If we add up the four values in the product column, we get

$$L=3.667+2.750-2.417-3.917=0.083 \tag{8.6.4}$$

This is the same value we got when we computed the difference between means previously (within rounding error). We call the value "L" for "linear combination."

Table $8.6.3$: *Coefficients for comparing low and high self-esteem*

Outcome	Esteem	Mean	Coeff	Product
Success	High Self-Esteem	7.333	0.5	3.667
	Low Self-Esteem	5.500	0.5	2.750
Failure	High Self-Esteem	4.833	-0.5	-2.417
	Low Self-Esteem	7.833	-0.5	-3.917

Now, the question is whether our value of L is significantly different from 0. The general formula for L is

$$L=\sum c_i M_i \tag{8.6.5}$$

where c_i is the i_{th} coefficient and M_i is the i_{th} mean. As shown above, $L=0.083$. The formula for testing L for significance is shown below:

$$t = \frac{L}{\sqrt{\frac{\sum c_i^2 MSE}{n}}} \tag{8.6.6}$$

In this example,

$$\sum c_i^2 = 0.5^2 + 0.5^2 + (-0.5)^2 + (-0.5)^2 = 1.00 \tag{8.6.7}$$

MSE is the mean of the variances. The four variances are shown in Table $8.6.4$. Their mean is 1.625. Therefore $MSE = 1.625$.

Table $8.6.4$: Variances of attributions of success or failure to oneself

Outcome	Esteem	Variance
Success	High Self-Esteem	1.867
	Low Self-Esteem	1.100
Failure	High Self-Esteem	2.167
	Low Self-Esteem	1.367

The value of n is the number of subjects in each group. Here $n = 6$.

Putting it all together,

$$t = \frac{0.083}{\sqrt{\frac{(1)(1.625)}{6}}} = 0.16 \tag{8.6.8}$$

We need to know the degrees of freedom in order to compute the probability value. The degrees of freedom is

$$df = N - k \tag{8.6.9}$$

where N is the total number of subjects (24) and k is the number of groups (4). Therefore, $df = 20$. Using the Online Calculator, we find that the two-tailed probability value is 0.874. Therefore, the difference between the "success" condition and the "failure" condition is not significant.

t distribution Calculator

A more interesting question about the results is whether the effect of outcome (success or failure) differs depending on the self-esteem of the subject. For example, success may make high-self-esteem subjects more likely to attribute the outcome to themselves, whereas success may make low-self-esteem subjects less likely to attribute the outcome to themselves.

To test this, we have to test a difference between differences. Specifically, is the difference between success and failure outcomes for the high-self-esteem subjects different from the difference between success and failure outcomes for the low-self-esteem subjects? The means shown in Table 5 show that this is the case. For the high-self-esteem subjects, the difference between the success and failure attribution scores is $7.333 - 4.833 = 2.500$. For low-self-esteem subjects, the difference is $5.500 - 7.833 = -2.333$. The difference between differences is $2.500 - (-2.333) = 4.833$.

The coefficients to test this difference between differences are shown in Table $8.6.5$.

Table $8.6.5$: Coefficients for testing difference between differences

Self-Esteem	Outcome	Mean	Coeff	Product
High	Success	7.333	1	7.333
	Failure	4.833	-1	-4.833
Low	Success	5.500	-1	-5.500
	Failure	7.833	1	7.83

If it is hard to see where these coefficients came from, consider that our difference between differences was computed this way:

$$\begin{aligned}(7.33 - 4.83) - (5.50 - 7.83) &= 7.33 - 4.83 - 5.50 + 7.83 \\ &= (1)7.33 + (-1)4.83 + (-1)5.50 + (1)7.83\end{aligned}$$

The values in parentheses are the coefficients.

To continue the calculations,

$$L = 4.83 \tag{8.6.10}$$

$$\sum c_i^2 = 1^2 + (-1)^2 + (-1)^2 + 1^2 = 4.00 \tag{8.6.11}$$

$$t = \frac{4.83}{\sqrt{\frac{(4)(1.625)}{6}}} = 4.64 \tag{8.6.12}$$

The two-tailed p value is 0.0002 Therefore, the difference between differences is highly significant.

In a later chapter on Analysis of Variance, you will see that comparisons such as this are testing what is called an interaction. In general, there is an interaction when the effect of one variable differs as a function of the level of another variable. In this example, the effect of the outcome variable is different depending on the subject's self-esteem. For the high-self-esteem subjects, success led to more self-attribution than did failure; for the low-self-esteem subjects, success led to less self-attribution than did failure.

Multiple Comparisons

The more comparisons you make, the greater your chance of a **Type I** error. It is useful to distinguish between two error rates:

1. the per-comparison error rate and
2. the familywise error rate

The per-comparison error rate is the probability of a **Type I** error for a particular comparison. The familywise error rate is the probability of making one or more **Type I** errors in a family or set of comparisons. In the attribution experiment discussed above, we computed two comparisons. If we use the 0.05 level for each comparison, then the per-comparison rate is simply 0.05. The familywise rate can be complex. Fortunately, there is a simple approximation that is fairly accurate when the number of comparisons is small. Defining α as the per-comparison error rate and c as the number of comparisons, the following inequality always holds true for the familywise error rate (FW):

$$FW \leq c\alpha \tag{8.6.13}$$

This inequality is called the **Bonferroni inequality**. In practice, FW can be approximated by $c\alpha$. This is a conservative approximation since FW can never be greater than $c\alpha$ and is generally less than $c\alpha$.

The **Bonferroni inequality** can be used to control the familywise error rate as follows: If you want the familywise error rate to be α, you use α/c as the per-comparison error rate. This correction, called the **Bonferroni correction**, will generally result in a familywise error rate less than α. Alternatively, you could multiply the probability value by c and use the original α level.

Should the familywise error rate be controlled? Unfortunately, there is no clear-cut answer to this question. The disadvantage of controlling the familywise error rate is that it makes it more difficult to obtain a significant result for any given comparison: The more comparisons you do, the lower the per-comparison rate must be and therefore the harder it is to reach significance. That is, the power is lower when you control the familywise error rate. The advantage is that you have a lower chance of making a **Type I** error.

One consideration is the definition of a family of comparisons. Let's say you conducted a study in which you were interested in whether there was a difference between male and female babies in the age at which they started crawling. After you finished analyzing the data, a colleague of yours had a totally different research question: Do babies who are born in the winter differ from those born in the summer in the age they start crawling? Should the familywise rate be controlled or should it be allowed to be greater than 0.05? Our view is that there is no reason you should be penalized (by lower power) just because your colleague used the same data to address a different research question. Therefore, the familywise error rate need not be controlled. Consider the two comparisons done on the attribution example at the beginning of this section: These comparisons are testing completely different hypotheses. Therefore, controlling the familywise rate is not necessary.

Now consider a study designed to investigate the relationship between various variables and the ability of subjects to predict the outcome of a coin flip. One comparison is between males and females; a second comparison is between those over 40 and those under 40; a third is between vegetarians and non-vegetarians; and a fourth is between firstborns and others. The question of whether these four comparisons are testing different hypotheses depends on your point of view. On the one hand, there is nothing about whether age makes a difference that is related to whether diet makes a difference. In that sense, the comparisons are addressing different hypotheses. On the other hand, the whole series of comparisons could be seen as addressing the general question of whether anything affects the ability to predict the outcome of a coin flip. If nothing does, then allowing the familywise rate to be high means that there is a high probability of reaching the wrong conclusion.

Orthogonal Comparisons

In the preceding sections, we talked about comparisons being independent. Independent comparisons are often called orthogonal comparisons. There is a simple test to determine whether two comparisons are orthogonal: If the sum of the products of the coefficients is 0, then the comparisons are orthogonal. Consider again the experiment on the attribution of success or failure. Table $8.6.6$ shows the coefficients previously presented in Table $8.6.3$ and in Table $8.6.5$. The column "$C1$" contains the coefficients from the comparison shown in Table $8.6.3$; the column "$C2$" contains the coefficients from the comparison shown in Table $8.6.5$. The column labeled "**Product**" is the product of these two columns. Note that the sum of the numbers in this column is 0. Therefore, the two comparisons are orthogonal.

Table $8.6.6$: *Coefficients for two orthogonal comparisons*

Outcome	Esteem	C1	C2	Product
Success	High Self-Esteem	0.5	1	0.5
	Low Self-Esteem	0.5	-1	-0.5
Failure	High Self-Esteem	-0.5	-1	0.5
	Low Self-Esteem	-0.5	1	-0.5

Table $8.6.7$ shows two comparisons that are not orthogonal. The first compares the high-self-esteem subjects to low-self-esteem subjects; the second considers only those in the success group and compares high-self-esteem subjects to low-self-esteem subjects. The failure group is ignored by using $0's$ as coefficients. Clearly the comparison of these two groups of subjects for the whole sample is not independent of the comparison of them for the success group only. You can see that the sum of the products of the coefficients is 0.5 and not 0.

Table $8.6.1$: *Coefficients for two non-orthogonal comparisons*

Outcome	Esteem	C1	C2	Product
Success	High Self-Esteem	0.5	0.5	0.25
	Low Self-Esteem	-0.5	-0.5	0.25
Failure	High Self-Esteem	0.5	0.0	0.0
	Low Self-Esteem	-0.5	0.0	0.0

8.7: Correlated Pairs

Learning Objectives

- Determine whether you have correlated pairs or independent groups
- Compute a t test for correlated pairs

Let's consider how to analyze the data from the "ADHD Treatment" case study. These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. In this section, we will be concerned only with testing the difference between the mean of the placebo ($D0$) condition and the mean of the highest dosage condition ($D60$). The first question is why the difference between means should not be tested using the procedure described in the section Difference Between Two Means (Independent Groups). The answer lies in the fact that in this experiment we do not have independent groups. The scores in the $D0$ condition are from the same subjects as the scores in the $D60$ condition. There is only one group of subjects, each subject being tested in both the $D0$ and $D60$ conditions.

Figure $8.7.1$ shows a scatter plot of the 60-mg scores ($D60$) as a function of the 0-mg scores ($D0$). It is clear that children who get more correct in the $D0$ condition tend to get more correct in the $D60$ condition. The correlation between the two conditions is high: $r = 0.80$. Clearly these two variables are not independent.

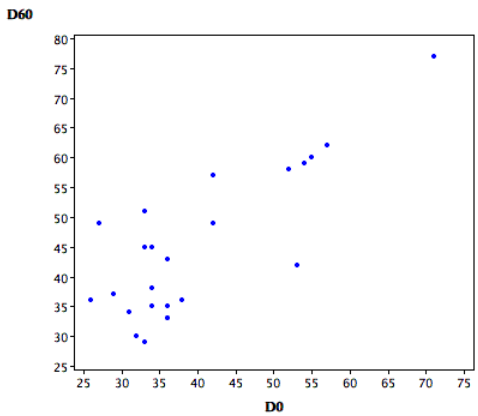


Figure $8.7.1$: Number of correct responses made in the 60-mg condition as a function of the number of correct responses in the 0-mg condition

Computations

You may recall that the method to test the difference between these means was presented in the section on "Testing a Single Mean." The computational procedure is to compute the difference between the $D60$ and the $D0$ conditions for each child and test whether the mean difference is significantly different from 0. The difference scores are shown in Table $8.7.1$. As shown in the section on testing a single mean, the mean difference score is 4.96 which is significantly different from 0: $t = 3.22,\ df = 23,\ p = 0.0038$. This t test has various names including "**correlated t test**" and "**related-pairs t test**".

In general, the correlated t test is computed by first computing the difference between the two scores for each subject. Then, a test of a single mean is computed on the mean of these difference scores.

Table $8.7.1$: DOG scores as a function of dosage

D0	D60	D60-D0
57	62	5
27	49	22
32	30	-2
31	34	3
34	38	4
38	36	-2

71	77	6
33	51	18
34	45	11
53	42	-11
36	43	7
42	57	15
26	36	10
52	58	6
36	35	-1
55	60	5
36	33	-3
42	49	7
36	33	-3
54	59	5
34	35	1
29	37	8
33	45	12
33	29	-4

If you had mistakenly used the method for an independent-groups t test with these data, you would have found that $t = 1.42$, $df = 46$, and $p = 0.15$. That is, the difference between means would not have been found to be statistically significant. This is a typical result: correlated t tests almost always have greater power than independent-groups t tests. This is because in correlated t tests, each difference score is a comparison of performance in one condition with the performance of that same subject in another condition. This makes each subject "their own control" and keeps differences between subjects from entering into the analysis. The result is that the standard error of the difference between means is smaller in the correlated t test and, since this term is in the denominator of the formula for t, results in a larger t.

Details about the Standard Error of the Difference between Means (Optional)

To see why the standard error of the difference between means is smaller in a correlated t test, consider the variance of difference scores. As shown in the section on the Variance Sum Law, the variance of the sum or difference of the two variables X and Y is:

$$S_{X \pm Y}^2 = S_X^2 + S_Y^2 \pm 2rS_XS_Y \tag{8.7.1}$$

Therefore, the variance of difference scores is the variance in the first condition (X) plus the variance in the second condition (Y) minus twice the product of

1. the correlation,
2. the standard deviation of X, and
3. the standard deviation of Y. For the current example, $r = 0.80$ and the variances and standard deviations are shown in Table 8.7.2.

Table 8.7.2: Variances and Standard Deviations

	D0	D60	D60 - D0
Variance	128.02	151.78	56.82

Sd	11.31	12.32	7.54

The variance of the difference scores of 56.82 can be computed as:

$$128.02 + 151.78 - (2)(0.80)(11.31)(12.32) \tag{8.7.2}$$

which is equal to 56.82 except for rounding error. Notice that the higher the correlation, the lower the standard error of the mean.

8.8: Correlated t Simulation

Learning Objectives

- State the effect of increasing r on power

Instructions

This simulation demonstrates the t test for correlated observations. When you click the "sample" button, 12 subjects with scores in two conditions are sampled from a population. Each population is normally distributed with a standard deviation of 4. The mean for condition 1 is 15 and the mean for condition 2 is 13. The initial population correlation between conditions is 0.5, although you can change that. The data from the sample of 12 subjects is graphed with a line connecting the two data points for a subject. The difference scores (Condition A - Condition B) are computed and graphed as well and shown in a table. The bottom of simulation displays the mean difference score, the standard deviation of the difference scores, the standard error of the mean difference score, the value of t, and the value of p.

Illustrated Instructions

Video Demo

The video begins by running simulations with the population correlation set to 0.5. More simulation are run with the correlation set to 0.1. As you watch the video and run simulations for yourself see if you can determine a correspondence between an aspect of the graph and the standard deviation of the difference scores. Specifically, look at the degree to which the lines diverge from being parallel and the standard deviation. What effect does increasing the correlation have on the parallel lines and the standard deviation? What effect does increasing the correlation have?

8.9: Specific Comparisons (Correlated Observations)

Learning Objectives

- Compute t for a comparison for repeated-measures data

In the "Weapons and Aggression" case study, subjects were asked to read words presented on a computer screen as quickly as they could. Some of the words were aggressive words such as injure or shatter. Others were control words such as relocate or consider. These two types of words were preceded by words that were either the names of weapons, such as shotgun or grenade, or non-weapon words, such as rabbit or fish. For each subject, the mean reading time across words was computed for these four conditions. The four conditions are labeled as shown in Table $8.9.1$. Table $8.9.2$ shows the data from five subjects.

Table $8.9.1$: Description of Conditions

Variable	Description
aw	The time in milliseconds (msec) to name an aggressive word following a weapon word prime.
an	The time in milliseconds (msec) to name an aggressive word following a non-weapon word prime.
cw	The time in milliseconds (msec) to name a control word following a weapon word prime.
cn	The time in milliseconds (msec) to name a control word following a non-weapon word prime.

Table $8.9.2$: Data from Five Subjects

Subject	aw	an	cw	cn
1	447	440	432	452
2	427	437	469	451
3	417	418	445	434
4	348	371	353	344
5	471	443	462	463

One question was whether reading times would be shorter when the preceding word was a weapon word (aw and cw conditions) than when it was a non-weapon word (an and cn conditions). In other words,

Is

$$L_1 = (an + cn) - (aw + cw) \tag{8.9.1}$$

greater than 0?

This is tested for significance by computing L_1 for each subject and then testing whether the mean value of L_1 is significantly different from 0. Table $8.9.3$ shows L_1 for the first five subjects. L_1 for Subject 1 was computed by:

$$L1 = (440 + 452) - (447 + 432) = 892 - 879 = 13 \tag{8.9.2}$$

Table $8.9.3$: L_1 for Five Subjects

Subject	aw	an	cw	cn	L_1
1	447	440	432	452	13
2	427	437	469	451	-8
3	417	418	445	434	-10

4	348	371	353	344	14
5	471	443	462	463	-27

Once L_1 is computed for each subject, the significance test described in the section "Testing a Single Mean" can be used. First we compute the mean and the standard error of the mean for L_1. There were 32 subjects in the experiment. Computing L_1 for the 32 subjects, we find that the mean and standard error of the mean are 5.875 and 4.2646, respectively. We then compute

$$t = \frac{M - \mu}{S_M} \tag{8.9.3}$$

where M is the sample mean, μ is the hypothesized value of the population mean (0 in this case), and s_M is the estimated standard error of the mean. The calculations show that $t = 1.378$. Since there were 32 subjects, the degrees of freedom is $32 - 1 = 31$. The t distribution calculator shows that the two-tailed probability is 0.178.

A more interesting question is whether the priming effect (the difference between words preceded by a non-weapon word and words preceded by a weapon word) is different for aggressive words than it is for non-aggressive words. That is, do weapon words prime aggressive words more than they prime non-aggressive words? The priming of aggressive words is $(an - aw)$. The priming of non-aggressive words is $(cn - cw)$. The comparison is the difference:

$$L_2 = (an - aw) - (cn - cw) \tag{8.9.4}$$

Table $8.9.4$ shows L_2 for five of the 32 subjects.

Table 8.9.4: L_2 for Five Subjects

Subject	aw	an	cw	cn	L_2
1	447	440	432	452	-27
2	427	437	469	451	28
3	417	418	445	434	12
4	348	371	353	344	32
5	471	443	462	463	-29

The mean and standard error of the mean for all 32 subjects are 8.4375 and 3.9128, respectively. Therefore, $t = 2.156$ and $p = 0.039$.

Multiple Comparisons

Issues associated with doing multiple comparisons are the same for related observations as they are for multiple comparisons among independent groups.

Orthogonal Comparisons

The most straightforward way to assess the degree of dependence between two comparisons is to correlate them directly. For the weapons and aggression data, the comparisons L_1 and L_2 are correlated 0.24. Of course, this is a sample correlation and only estimates what the correlation would be if L_1 and L_2 were correlated in the population. Although mathematically possible, orthogonal comparisons with correlated observations are very rare.

8.10: Pairwise (Correlated)

Learning Objectives

- Compute the Bonferroni correction
- Calculate pairwise comparisons using the Bonferroni correction

In the section on all pairwise comparisons among independent groups, the Tukey HSD test was the recommended procedure. However, when you have one group with several scores from the same subjects, the Tukey test makes an assumption that is unlikely to hold: The variance of difference scores is the same for all pairwise differences between means.

The standard practice for pairwise comparisons with correlated observations is to compare each pair of means using the method outlined in the section "Difference Between Two Means (Correlated Pairs)" with the addition of the Bonferroni correction described in the section "Specific Comparisons." For example, suppose you were going to do all pairwise comparisons among four means and hold the familywise error rate at 0.05. Since there are six possible pairwise comparisons among four means, you would use $0.05/6 = 0.0083$for the per-comparison error rate.

As an example, consider the case study "Stroop Interference." There were three tasks, each performed by 47 subjects. In the "words" task, subjects read the names of 60 color words written in black ink; in the "color" task, subjects named the colors of 60 rectangles; in the "interference" task, subjects named the ink color of 60 conflicting color words. The times to read the stimuli were recorded. In order to compute all pairwise comparisons, the difference in times for each pair of conditions for each subject is calculated. Table $8.10.1$shows these scores for five of the 47 subjects.

Table 8.10.1: Pairwise Differences

W-C	W-I	C-I
-3	-24	-21
2	-41	-43
-1	-18	-17
-4	-23	-19
-2	-17	-15

Table $8.10.2$shows data for all 47 subjects.

Table* $8.10.2$*: *Pairwise Differences for* 47 *subjects*

W-C	W-I	C-I
-3	-24	-21
2	-41	-43
-1	-18	-17
-4	-23	-19
-2	-17	-15
-3	-15	-12
-3	-28	-25
-3	-36	-33

-3	-17	-14
-2	-10	-8
-1	-11	-10
-1	-10	-9
-3	-26	-23
0	-4	-4
-4	-28	-24
-5	-19	-14
-5	-18	-13
-8	-23	-15
-7	-22	-15
-3	-28	-25
-9	-30	-21
-5	-21	-16
-7	-23	-16
-9	-21	-12
-7	-35	-28
-4	-27	-23
-4	-25	-21
-2	-16	-14
-3	-14	-11
-5	-14	-9
-3	-14	-11
-5	-15	-10
-12	-25	-13
-1	-9	-8
-2	-13	-11
-8	-17	-9

3	-13	-16
-6	-33	-27
-3	-12	-9
-7	-19	-12
-8	-19	-11
-6	-31	-25
-1	-19	-18
-5	-13	-8
-3	-18	-15
-9	-28	-19
-5	-22	-17

The means, standard deviations (Sd), and standard error of the mean (Sem), t, and p for all 47 subjects are shown in Table $8.10.3$. The $t's$ are computed by dividing the means by the standard errors of the mean. Since there are 47 subjects, the degrees of freedom is 46. Notice how different the standard deviations are. For the Tukey test to be valid, all population values of the standard deviation would have to be the same.

Table $8.10.3$: Pairwise Comparisons

Comparison	Mean	Sd	Sem	t	p
W-C	-4.15	2.99	0.44	-9.53	<0.001
W-I	-20.51	7.84	1.14	-17.93	<0.001
C-I	-16.36	7.47	1.09	-15.02	<0.001

Using the **Bonferroni correction** for three comparisons, the p value has to be below $0.05/3 = 0.0167$ for an effect to be significant at the 0.05 level. For these data, all p values are far below that, and therefore all pairwise differences are significant.

8.11: Statistical Literacy

Learning Objectives

- Effectiveness of Surgery for Weight Loss

Research on the effectiveness of surgery for weight loss reported here found that "The surgery was associated with significantly greater weight loss [than the control group who dieted] through 2 years (61.3 versus 11.2 pounds, $P < 0.001$)."

Example $8.11.1$: what do you think

What test could have been used and how would it have been computed?

Solution

For each subject a difference score between their initial weight and final weight could be computed. A t test of whether the mean difference score differs significantly from 0 could then be computed. The mean difference score will equal the difference between the mean weight losses of the two groups ($61.3 - 11.2 = 50.1$).

8.E: Tests of Means (Exercises)

General Questions

Q1

The scores of a random sample of 8 students on a physics test are as follows: $60, 62, 67, 69, 70, 72, 75, 78$

a. Test to see if the sample mean is significantly different from 65 at the 0.05 level. Report the t and p values.

b. The researcher realizes that she accidentally recorded the score that should have been 76 as 67. Are these corrected scores significantly different from 65 at the 0.05 level? (relevant section)

Q2

A (hypothetical) experiment is conducted on the effect of alcohol on perceptual motor ability. Ten subjects are each tested twice, once after having two drinks and once after having two glasses of water. The two tests were on two different days to give the alcohol a chance to wear off. Half of the subjects were given alcohol first and half were given water first. The scores of the 10 subjects are shown below. The first number for each subject is their performance in the "water" condition. Higher scores reflect better performance. Test to see if alcohol had a significant effect. Report the t and p values. (relevant section)

water	alcohol
16	13
15	13
11	10
20	18
19	17
14	11
13	10
15	15
14	11
16	16

Q3

The scores on a (hypothetical) vocabulary test of a group of 20 year olds and a group of 60 year olds are shown below.

20 yr olds	60 yr olds
27	26
26	29
21	29
24	29
15	27
18	16
17	20
12	27
13	

a. Test the mean difference for significance using the 0.05 level. (relevant section).
b. List the assumptions made in computing your answer.(relevant section)

Q4

The sampling distribution of a statistic is normally distributed with an estimated standard error of 12, $(df = 20)$.

a. What is the probability that you would have gotten a mean of 107 (or more extreme) if the population parameter were 100? Is this probability significant at the 0.05 level (two-tailed)?
b. What is the probability that you would have gotten a mean of 95 or less (one-tailed)? Is this probability significant at the 0.05 level? You may want to use the t Distribution calculator for this problem. (relevant section)

Q5

How do you decide whether to use an independent groups t test or a correlated t test (test of dependent means)? (relevant section & relevant section)

Q6

An experiment compared the ability of three groups of subjects to remember briefly-presented chess positions. The data are shown below.

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

a. Using the Tukey HSD procedure, determine which groups are significantly different from each other at the 0.05 level. (relevant section)
b. Now compare each pair of groups using t-tests. Make sure to control for the familywise error rate (at 0.05) by using the **Bonferroni correction**. Specify the alpha level you used.

Q7

Below are data showing the results of six subjects on a memory test. The three scores per subject are their scores on three trials (a, b, and c) of a memory task. Are the subjects getting better each trial? Test the linear effect of trial for the data.

a	b	c
4	6	7
3	7	8
2	8	5
1	4	7
4	6	9

2	4	2

a. Compute L for each subject using the contrast weights -1, 0, and 1. That is, compute $(-1)(a)+(0)(b)+(1)(c)$ for each subject.
b. Compute a one-sample t-test on this column (with the L values for each subject) you created. (relevant section)

Q8

Participants threw darts at a target. In one condition, they used their preferred hand; in the other condition, they used their other hand. All subjects performed in both conditions (the order of conditions was counterbalanced). Their scores are shown below.

Preferred	Non-preferred
12	7
7	9
11	8
13	10
10	9

a. Which kind of t-test should be used?
b. Calculate the two-tailed t and p values using this t test.
c. Calculate the one-tailed t and p values using this t test.

Q9

Assume the data in the previous problem were collected using two different groups of subjects: One group used their preferred hand and the other group used their non-preferred hand. Analyze the data and compare the results to those for the previous problem (relevant section)

Q10

You have 4 means, and you want to compare each mean to every other mean.

a. How many tests total are you going to compute?
b. What would be the chance of making at least one **Type I** error if the **Type I** error for each test was 0.05 and the tests were independent? (relevant section & relevant section)
c. Are the tests independent and how does independence/non-independence affect the probability in (b).

Q11

In an experiment, participants were divided into 4 groups. There were 20 participants in each group, so the degrees of freedom (error) for this study was $80-4=76$. Tukey's HSD test was performed on the data.

a. Calculate the p value for each pair based on the Q value given below. You will want to use the Studentized Range Calculator.
b. Which differences are significant at the 0.05 level? (relevant section)

Comparison of Groups	Q
A - B	3.4
A - C	3.8
A - D	4.3
B - C	1.7
B - D	3.9
C - D	3.7

Q12

If you have 5 groups in your study, why shouldn't you just compute a t test of each group mean with each other group mean? (relevant section)

Q13

You are conducting a study to see if students do better when they study all at once or in intervals. One group of 12 participants took a test after studying for one hour continuously. The other group of 12 participants took a test after studying for three twenty minute sessions. The first group had a mean score of 75 and a variance of 120. The second group had a mean score of 86 and a variance of 100.

a. What is the calculated t value? Are the mean test scores of these two groups significantly different at the 0.05 level?
b. What would the t value be if there were only 6 participants in each group? Would the scores be significant at the 0.05 level?

Q14

A new test was designed to have a mean of 80 and a standard deviation of 10. A random sample of 20 students at your school take the test, and the mean score turns out to be 85. Does this score differ significantly from 80? To answer this problem, you may want to use the Normal Distribution Calculator.(relevant section)

Q15

You perform a one-sample t test and calculate a t statistic of 3.0. The mean of your sample was 1.3 and the standard deviation was 2.6. How many participants were used in this study? (relevant section)

Q16

True/false: The contrasts $(-3, 111)$ and $(0, 0, -1, 1)$ are orthogonal. (relevant section)

Q17

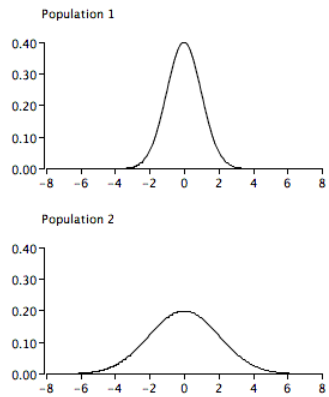
True/false: If you are making 4 comparisons between means, then based on the **Bonferroni correction**, you should use an alpha level of 0.01 for each test. (relevant section)

Q18

True/false: Correlated t tests almost always have greater power than independent t tests. (relevant section)

Q19

True/false:The graph below represents a violation of the homogeneity of variance assumption. (relevant section)



Q20

True/false: When you are conducting a one-sample t test and you know the population standard deviation, you look up the critical t value in the table based on the degrees of freedom. (relevant section)

Questions from Case Studies

The following questions use data from the Angry Moods (AM) case study.

Q21

(AM#17) Do athletes or non-athletes calm down more when angry? Conduct a t test to see if the difference between groups in Control-In scores is statistically significant.

Q22

Do people in general have a higher Anger-Out or Anger-In score? Conduct a t test on the difference between means of these two scores. Are these two means independent or dependent? (relevant section)

The following questions use data from the Smiles and Leniency (SL) case study.

Q23

Compare each mean to the neutral mean. Be sure to control for the familywise error rate. (relevant section)

Q24

Does a "felt smile" lead to more leniency than other types of smiles?

a. Calculate L (the linear combination) using the following contrast weights $false: -1,\ felt: 2,\ miserable: -1,\ neutral: 0$.
b. Perform a significance test on this value of L. (relevant section)

The following questions are from the Animal Research (AR) case study.

Q25

(AR#8) Conduct an independent samples t test comparing males to females on the belief that animal research is necessary. (relevant section)

Q26

(AR#9) Based on the t test you conducted in the previous problem, are you able to reject the null hypothesis if $alpha = 0.05$? What about if $alpha = 0.1$? (relevant section)

Q27

(AR#10) Is there any evidence that the t test assumption of homogeneity of variance is violated in the t test you computed in #25? (relevant section)

The following questions use data from the ADHD Treatment (AT) case study.

Q28

Compare each dosage with the dosage below it (compare $d0$ and $d15$, $d15$ and $d30$, and $d30$ and $d60$). Remember that the patients completed the task after every dosage.

a. If the familywise error rate is 0.05, what is the alpha level you will use for each comparison when doing the Bonferroni correction?
b. Which differences are significant at this level? (relevant section)

Q29

Does performance increase linearly with dosage?

a. Plot a line graph of this data.
b. Compute L for each patient. To do this, create a new variable where you multiply the following coefficients by their corresponding dosages and then sum up the total: $(-3)d0 + (-1)d15 + (1)d30 + (3)d60$ (see #8). What is the mean of L?
c. Perform a significance test on L. Compute the 95% confidence interval for L. (relevant section)

Select Answers

S1

a. $t(7) = 1.91$

S4

b. 0.035

S7

b. two-tailed $p = 0.0088$

S8

b. $p = 0.1662$

S11

a. $A - B : p = 0.085$

S13

a. $t(22) = 2.57$

S23

$t(76) = 3.04$

S25

a. $p = 0.0745$

S29

c. $p = 0.0006$

CHAPTER OVERVIEW

9: Describing Bivariate Data

Probability is an important and complex field of study. Fortunately, only a few basic issues in probability theory are essential for understanding statistics at the level covered in this book. These basic issues are covered in this chapter. The introductory section discusses the definitions of probability. This is not as simple as it may seem. The section on basic concepts covers how to compute probabilities in a variety of simple situations. The Gambler's Fallacy Simulation provides an opportunity to explore this fallacy by simulation. The Birthday Demonstration illustrates the probability of finding two or more people with the same birthday. The Binomial Demonstration shows the binomial distribution for different parameters. The section on base rates discusses an important but often-ignored factor in determining probabilities. It also presents Bayes' Theorem. The Bayes' Theorem Demonstration shows how a tree diagram and Bayes' Theorem result in the same answer. Finally, the Monty Hall Demonstration lets you play a game with a very counterintuitive result.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

9.1: Introduction to Bivariate Data

Learning Objectives

- Define "bivariate data"
- Define "scatter plot"
- Distinguish between a linear and a nonlinear relationship
- Identify positive and negative associations from a scatter plot

Measures of central tendency, variability, and spread summarize a single variable by providing important information about its distribution. Often, more than one variable is collected on each individual. For example, in large health studies of populations it is common to obtain variables such as age, sex, height, weight, blood pressure, and total cholesterol on each individual. Economic studies may be interested in, among other things, personal income and years of education. As a third example, most university admissions committees ask for an applicant's high school grade point average and standardized admission test scores (e.g., SAT). In this chapter we consider bivariate data, which for now consists of two quantitative variables for each individual. Our first interest is in summarizing such data in a way that is analogous to summarizing univariate (single variable) data.

By way of illustration, let's consider something with which we are all familiar: age. Let’s begin by asking if people tend to marry other people of about the same age. Our experience tells us "yes," but how good is the correspondence? One way to address the question is to look at pairs of ages for a sample of married couples. Table $9.1.1$ below shows the ages of 10 married couples. Going across the columns we see that, yes, husbands and wives tend to be of about the same age, with men having a tendency to be slightly older than their wives. This is no big surprise, but at least the data bear out our experiences, which is not always the case.

Table $9.1.1$: *Sample of spousal ages of* 10 *White American Couples*

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

The pairs of ages in Table $9.1.1$ are from a dataset consisting of 282 pairs of spousal ages, too many to make sense of from a table. What we need is a way to summarize the 282 pairs of ages. We know that each variable can be summarized by a histogram (see Figure $9.1.1$) and by a mean and standard deviation (See Table $9.1.2$).

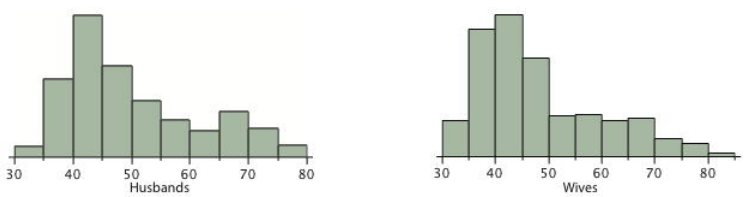


Figure $9.1.1$: Histograms of spousal ages

Table $9.1.2$: Means and standard deviations of spousal ages

	Mean	Standard Deviation
Husbands	49	11
Wives	47	11

Each distribution is fairly skewed with a long right tail. From Table $9.1.1$ we see that not all husbands are older than their wives and it is important to see that this fact is lost when we separate the variables. That is, even though we provide summary statistics on each variable, the pairing within couple is lost by separating the variables. We cannot say, for example, based on the means alone what percentage of couples has younger husbands than wives. We have to count across pairs to find this out. Only by maintaining the pairing can meaningful answers be found about couples per se. Another example of information not available from the separate descriptions of husbands and wives' ages is the mean age of husbands with wives of a certain age. For instance, what is the average age of husbands with 45-year-old wives? Finally, we do not know the relationship between the husband's age and the wife's age.

We can learn much more by displaying the bivariate data in a graphical form that maintains the pairing. Figure $9.1.2$ shows a scatter plot of the paired ages. The x-axis represents the age of the husband and the y-axis the age of the wife.

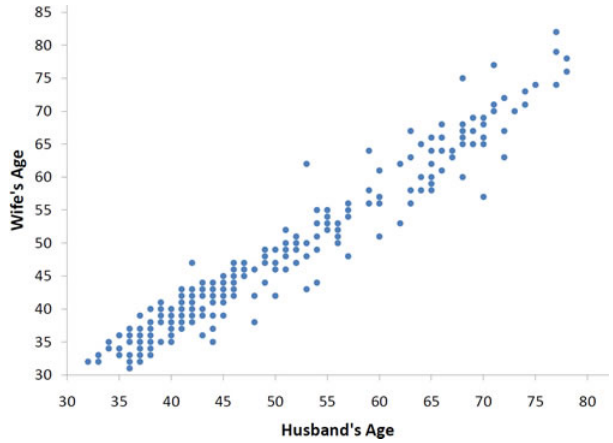


Figure $9.1.2$: Scatter plot showing wife's age as a function of husband's age

There are two important characteristics of the data revealed by Figure $9.1.2$. First, it is clear that there is a strong relationship between the husband's age and the wife's age: the older the husband, the older the wife. When one variable (Y) increases with the second variable (X), we say that X and Y have a positive association. Conversely, when Y decreases as X increases, we say that they have a negative association.

Second, the points cluster along a straight line. When this occurs, the relationship is called a linear relationship.

Figure $9.1.3$ shows a scatter plot of Arm Strength and Grip Strength from 149 individuals working in physically demanding jobs including electricians, construction and maintenance workers, and auto mechanics. Not surprisingly, the stronger someone's grip, the stronger their arm tends to be. There is therefore a positive association between these variables. Although the points cluster along a line, they are not clustered quite as closely as they are for the scatter plot of spousal age.

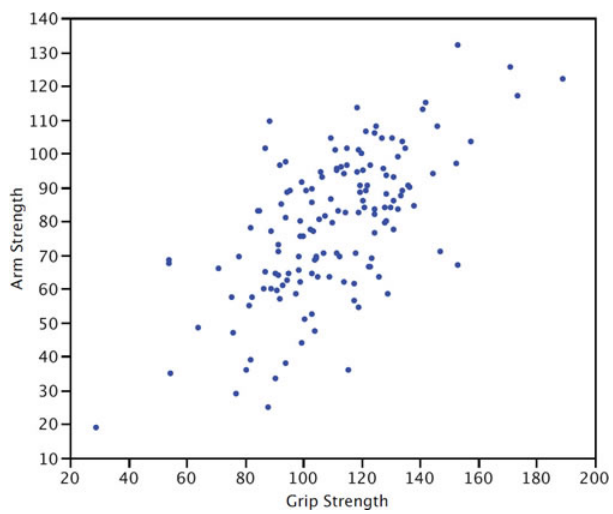


Figure $9.1.3$: Scatter plot of Grip Strength and Arm Strength

Not all scatter plots show linear relationships. Figure $9.1.4$ shows the results of an experiment conducted by Galileo on projectile motion. In the experiment, Galileo rolled balls down an incline and measured how far they traveled as a function of the release height. It is clear from Figure $9.1.4$ that the relationship between "Release Height" and "Distance Traveled" is not described well by a straight line: If you drew a line connecting the lowest point and the highest point, all of the remaining points would be above the line. The data are better fit by a parabola.

D. Dickey and T. Arnold's description of the study including a movie

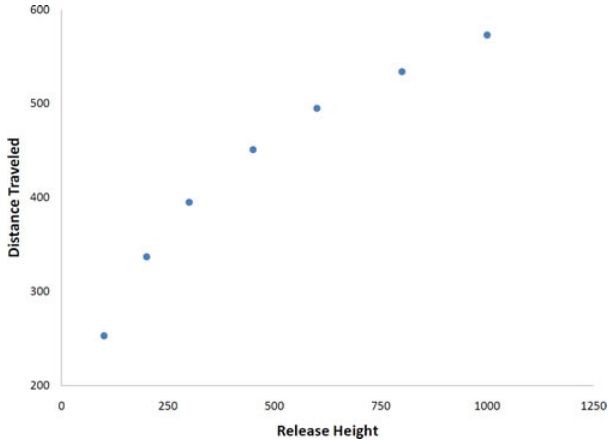


Figure 9.1.4: Galileo's data showing a non-linear relationship

Scatter plots that show linear relationships between variables can differ in several ways including the slope of the line about which they cluster and how tightly the points cluster about the line. A statistical measure of the strength of the relationship between two quantitative variables that takes these factors into account is the subject of the section "Values of Pearson's Correlation."

Contributor

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
- Rudy Guerra and David M. Lane

9.2: Values of the Pearson Correlation

Learning Objectives

- Describe what Pearson's correlation measures
- Give the symbols for Pearson's correlation in the sample and in the population
- State the possible range for Pearson's correlation
- Identify a perfect linear relationship

The Pearson product-moment correlation coefficient is a measure of the strength of the *linear relationship* between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Because we will be dealing almost exclusively with samples, we will use r to represent Pearson's correlation unless otherwise noted.

Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables. Figure $9.2.1$ shows a scatter plot for which $r = 1$.

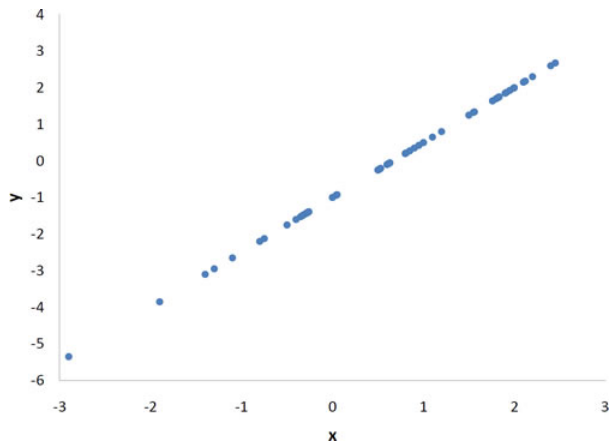


Figure $9.2.1$: *A perfect positive linear relationship*, $r = 1$

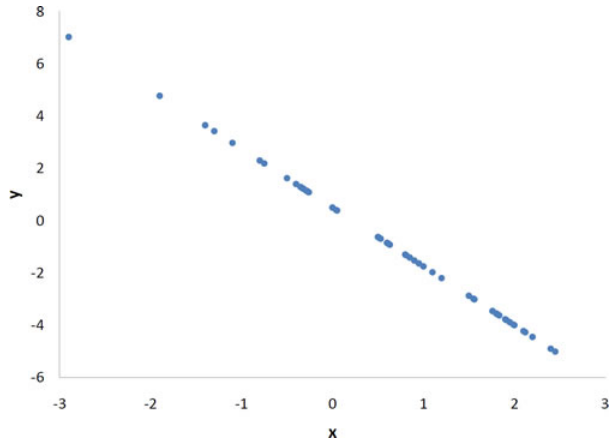


Figure $9.2.2$: *A perfect negative linear relationship*, $r = -1$

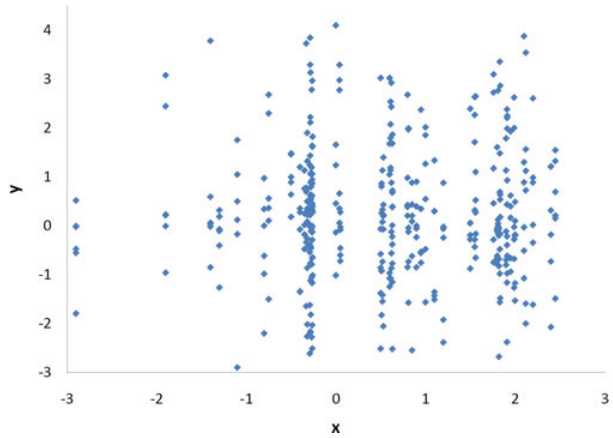


Figure $9.2.3$: *A scatter plot for which* $r = 0$. *Notice that there is no relationship between* X *and* Y.

With real data, you would not expect to get values of r of exactly $-1, 0,$ *or* 1. The data for spousal ages shown in Figure $9.2.4$ and described in the introductory section has an r of 0.97.

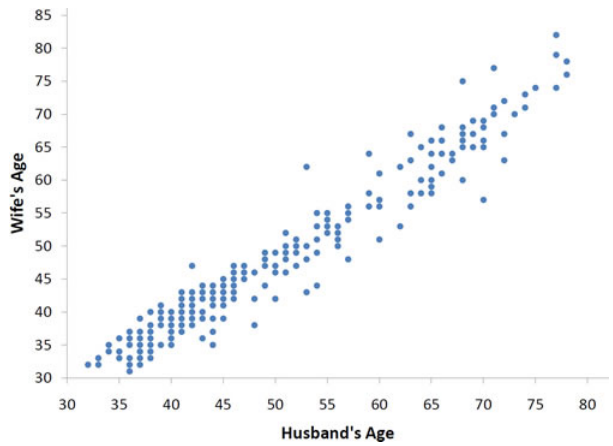


Figure $9.2.4$: *Scatter plot of spousal ages,* $r = 0.97$

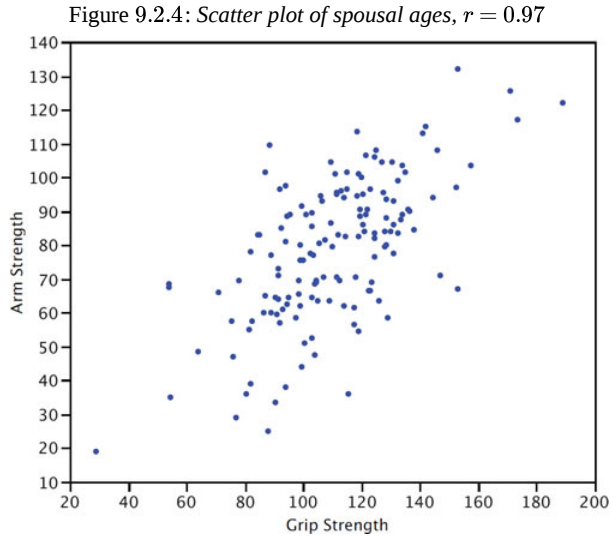


Figure $9.2.5$: *Scatter plot of Grip Strength and Arm Strength,* $r = 0.63$

The relationship between grip strength and arm strength depicted in Figure $9.2.5$ (also described in the introductory section) is 0.63.

9.3: Guessing Correlations

Learning Objectives

- Learn how to estimate Pearson's correlation coefficient by viewing scatter plots

Instructions

This demonstration allows you to learn about Pearson's correlation by viewing scatter plots with different values of Pearson's r. In each case, you will have an opportunity to guess the correlation. With a little practice, you should get pretty good at it.

Illustrated Instructions

The demonstration begins by displaying a scatterplot (shown below) and you guess what the correlation of the data in the scatterplot is by selecting one of five options below the graph.

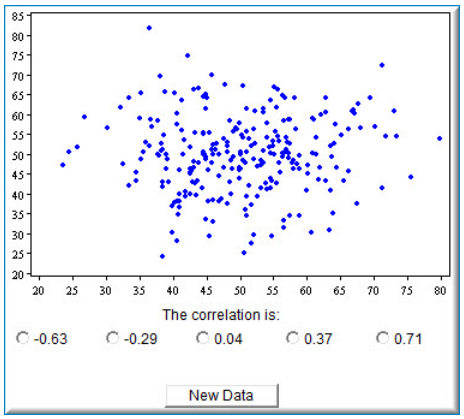


Figure 9.3.1: Beginning Scatterplot

As can be seen in screenshots below, once you select an answer you will be prompted whether the answer is correct or incorrect. Click the "New Data" to get a new dataset.

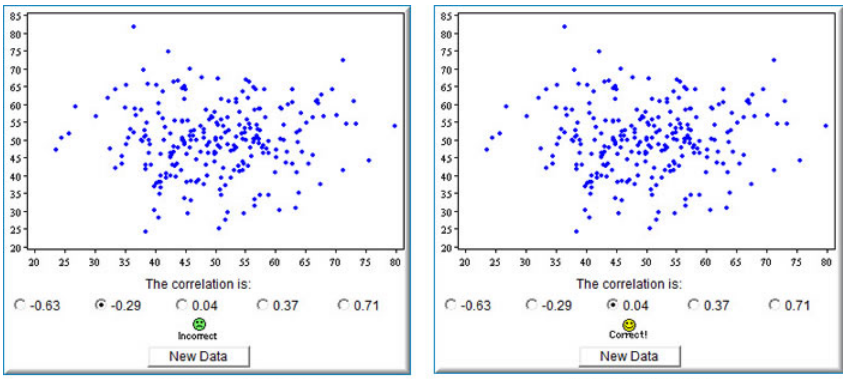


Figure 9.3.2: Scatterplot after selecting an answer

9.4: Properties of r

Learning Objectives

- State the range of values for Pearson's correlation
- State the values that represent perfect linear relationships
- State the relationship between the correlation of Y with X and the correlation of X with Y
- State the effect of linear transformations on Pearson's correlation

A basic property of Pearson's r is that its possible range is from -1 to 1. A correlation of -1 means a perfect negative linear relationship, a correlation of 0 means no linear relationship, and a correlation of 1 means a perfect positive linear relationship.

Pearson's correlation is symmetric in the sense that the correlation of X with Y is the same as the correlation of Y with X. For example, the correlation of Weight with Height is the same as the correlation of Height with Weight.

A critical property of Pearson's r is that it is unaffected by linear transformations. This means that multiplying a variable by a constant and/or adding a constant does not change the correlation of that variable with other variables. For instance, the correlation of Weight and Height does not depend on whether Height is measured in inches, feet, or even miles. Similarly, adding five points to every student's test score would not change the correlation of the test score with other variables such as GPA.

9.5: Computing r

Learning Objectives

- Define X and x
- State why $\sum xy = 0$ when there is no relationship
- Calculate r

There are several formulas that can be used to compute Pearson's correlation. Some formulas make more conceptual sense whereas others are easier to actually compute. We are going to begin with a formula that makes more conceptual sense.

We are going to compute the correlation between the variables X and Y shown in Table $9.5.1$. We begin by computing the mean for X and subtracting this mean from all values of X. The new variable is called "x". The variable "y" is computed similarly. The variables x and y are said to be deviation scores because each score is a deviation from the mean. Notice that the means of x and y are both 0. Next we create a new column by multiplying x and y.

Before proceeding with the calculations, let's consider why the sum of the xy column reveals the relationship between X and Y. If there were no relationship between X and Y, then positive values of x would be just as likely to be paired with negative values of y as with positive values. This would make negative values of xy as likely as positive values and the sum would be small. On the other hand, consider Table 1 in which high values of X are associated with high values of Y and low values of X are associated with low values of Y. You can see that positive values of x are associated with positive values of y and negative values of x are associated with negative values of y. In all cases, the product of x and y is positive, resulting in a high total for the xy column. Finally, if there were a negative relationship then positive values of x would be associated with negative values of y and negative values of x would be associated with positive values of y. This would lead to negative values for xy.

Table $9.5.1$: Calculation of r

	X	Y	x	y	xy	x^2	y^2
	1	4	-3	-5	15	9	25
	3	6	-1	-3	3	1	9
	5	10	1	1	1	1	1
	5	12	1	3	3	1	9
	6	13	2	4	8	4	16
Total	20	45	0	0	30	16	60
Mean	4	9	0	0	6		

Pearson's r is designed so that the correlation between height and weight is the same whether height is measured in inches or in feet. To achieve this property, Pearson's correlation is computed by dividing the sum of the xy column ($\sum xy$) by the square root of the product of the sum of the x^2 column ($\sum x^2$) and the sum of the y^2 column ($\sum y^2$). The resulting formula is:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \tag{9.5.1}$$

and therefore

$$r = \frac{30}{\sqrt{(16)(60)}} = \frac{30}{\sqrt{960}} = \frac{30}{30.984} = 0.968 \tag{9.5.2}$$

An alternative computational formula that avoids the step of computing deviation scores is:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)}\sqrt{\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}} \tag{9.5.3}$$

9.6: Restriction of Range Demo

Learning Objectives

- State the effect of choosing the upper and lower extremes on the correlation

Instructions

This demonstration illustrates the effect of restricting the range of scores on the the correlaton between variables. When the demonstration begins, the dataset "Strength" is displayed. The X-axis contains scores on a measure of grip strength; the Y-axis contains scores on a measure of arm strength. Notice that there is a strong relationship. Also note that the two scatterplots are identical and both show that for the entire dataset of 147 subjects, the correlaton is 0.63. The two vertical bars on the upper graph can be used to select a subset of scores to be plotted on the lower graph. If you click on the left-hand bar and drag it to the right, scores to the left of the bar will not be included on the lower graph. The correlaton shown for the lower graph shows only the correlation for the data in the subset. Similarly, if you drag the right-hand bar to the left, scores to the right of the bar will be excluded. If you click the "Data outside bars" button, the lower graph will, as you might expect, only include scores to the left of the left-hand bar and to the right of the right-hand bar.

Illustrated Instructions

The restriction of range simulation displays data from 1 of 4 datasets. You can change datasets by selecting one from the drop down menu at the top of the simulation. The top scatterplot displays all the data from the selected dataset.

Notice the two vertical bars on either side of this graph. You can "drag" these bars and restrict the range of data from the dataset that is displayed in the bottom scatterplot. You can include data on the inside or outside of the bars by selecting one of the options right above the graph.

For some browsers you will not see the bars move as you move them. They will move when you release the mouse button.

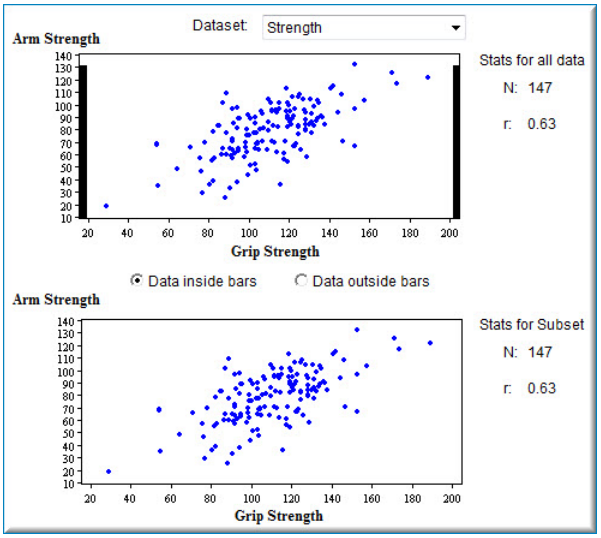


Figure 9.6.1: Restriction of Range demonstration

9.7: Variance Sum Law II - Correlated Variables

Learning Objectives

- State the variance sum law when X and Y are not assumed to be independent
- Compute the variance of the sum of two variables if the variance of each and their correlation is known
- Compute the variance of the difference between two variables if the variance of each and their correlation is known

Recall that when the variables X and Y are independent, the variance of the sum or difference between X and Y can be written as follows:

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y \tag{9.7.1}$$

which is read: "The variance of X plus or minus Y is equal to the variance of X plus the variance of Y."

When X and Y are correlated, the following formula should be used:

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho\sigma_X\sigma_Y \tag{9.7.2}$$

where ρ is the correlation between X and Y in the **population**.

Example 9.7.1

If the variance of verbal SAT were $10,000$, the variance of quantitative SAT were $11,000$ and the correlation between these two tests were 0.50, what is the variance of total SAT (verbal + quantitative) and the difference (verbal - quantitative)?

Solution

Since the two variables are correlated, we use Equation 9.7.2 instead of Equation 9.7.1 for uncorrelated (independent) variables. Hence, the variance of the sum is

$$\sigma^2_{verbal+quant} = 10,000 + 11,000 + 2 \times 0.5 \times \sqrt{10,000} \times \sqrt{11,000} \tag{9.7.3}$$

which is equal to $31,488$. The variance of the difference is also determined by Equation 9.7.2:

$$\sigma^2_{verbal-quant} = 10,000 + 11,000 - 2 \times 0.5 \times \sqrt{10,000} \times \sqrt{11,000} \tag{9.7.4}$$

which is equal to $10,512$.

If the variances and the correlation are computed in a sample, then the following notation is used to express the variance sum law:

$$s^2_{X \pm Y} = s^2_X + s^2_Y \pm 2r\, s_X\, s_Y \tag{9.7.5}$$

9.8: Statistical Literacy

Learning Objectives

- Relationship between age and sleep

Age and Sleep

The graph below showing the relationship between age and sleep is based on a graph that appears on this web page.

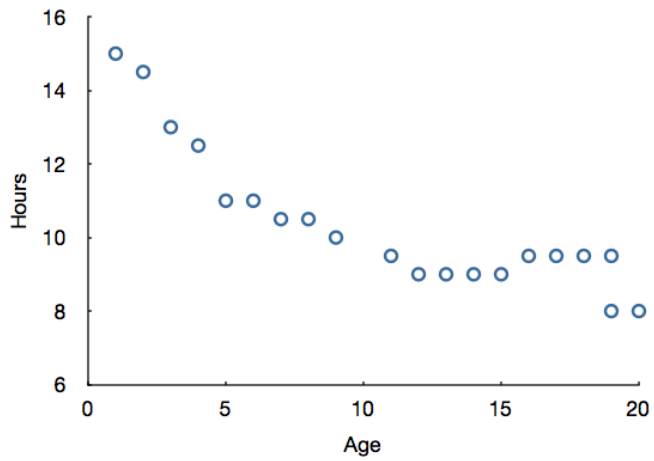


Figure 9.8.1: Graph showing relationship between age and sleep

Example 9.8.1

Based on the graph in Figure 9.8.1 Why might Pearson's correlation not be a good way to describe the relationship?

Solution

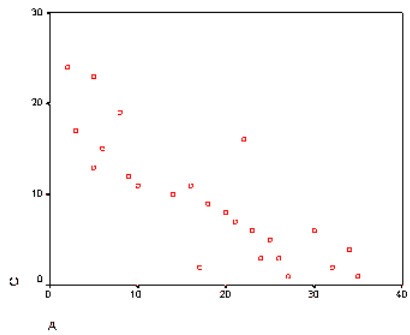
Pearson's correlation measures the strength of the linear relationship between two variables. The relationship here is not linear. As age increases, hours slept decreases rapidly at first but then levels off.

9.E: Describing Bivariate Data (Exercises)

General Questions

Q1

Describe the relationship between variables A and C. Think of things these variables could represent in real life. (relevant section)



Q2

Make up a data set with 10 numbers that has a positive correlation. (relevant section & relevant section)

Q3

Make up a data set with 10 numbers that has a negative correlation. (relevant section & relevant section)

Q4

If the correlation between weight (in pounds) and height (in feet) is 0.58, find: (relevant section)

a. the correlation between weight (in pounds) and height (in yards)
b. the correlation between weight (in kilograms) and height (in meters)

Q5

Would you expect the correlation between High School GPA and College GPA to be higher when taken from your entire high school class or when taken from only the top 20 students? Why? (relevant section)

Q6

For a certain class, the relationship between the amount of time spent studying and the test grade earned was examined. It was determined that as the amount of time they studied increased, so did their grades. Is this a positive or negative association? (relevant section)

Q7

For this same class, the relationship between the amount of time spent studying and the amount of time spent socializing per week was also examined. It was determined that the more hours they spent studying, the fewer hours they spent socializing. Is this a positive or negative association? (relevant section)

Q8

For the following data: (relevant section)

A	B
2	8

5	5
6	2
8	4
9	1

a. Find the deviation scores for Variable A that correspond to the raw scores of 2 and 8.
b. Find the deviation scores for Variable B that correspond to the raw scores of 5 and 4.
c. Just from looking at these scores, do you think these variable A and B are positively or negatively correlated? Why?
d. Now calculate the correlation. Were you right?

Q9

Students took two parts of a test, each worth 50 points. Part A has a variance of 25, and Part B has a variance of 36. The correlation between the test scores is 0.8.

a. If the teacher adds the grades of the two parts together to form a final test grade, what would the variance of the final test grades be?
b. What would the variance of Part A - Part B be? (relevant section)

Q10

True/False: The correlation in real life between height and weight is $r = 1$. (relevant section)

Q11

True/False: It is possible for variables to have $r = 0$ but still have a strong association. (relevant section & relevant section)

Q12

True/False: Two variables with a correlation of 0.3 have a stronger linear relationship than two variables with a correlation of -0.7. (relevant section)

Q13

True/False: After polling a certain group of people, researchers found a 0.5 correlation between the number of car accidents per year and the driver's age. This means that older people get in more accidents. (relevant section)

Q14

True/False: The correlation between R and T is the same as the correlation between T and R. (relevant section)

Q15

True/False: To examine bivariate data graphically, the best choice is two side by side histograms. (relevant section)

Q16

True/False: A correlation of $r = 1.2$ is not possible. (relevant section)

Questions from Case Studies

The following questions are from the Angry Moods (AM) case study.

Q17

(AM#21) What is the correlation between the Control-In and Control-Out scores? (relevant section)

Q18

(AM#22) Would you expect the correlation between the Anger-Out and Control-Out scores to be positive or negative? Compute this correlation. (relevant section & relevant section)

The following question is from the Flatulence (F) case study.

Q19

(F#4) Is there a relationship between the number of male siblings and embarrassment in front of romantic interests? Create a scatterplot and compute r. (relevant section & relevant section)

The following questions are from the Stroop (S) case study.

Q20

(S#6) Create a scatterplot showing "colors" on the Y-axis and "words" on the X-axis. (relevant section)

Q21

(S#7) Compute the correlation between "colors" and "words." (relevant section)

Q22

Sort the data by color-naming time. Choose only the 20 fastest color-namers and create a scatterplot. (relevant section)

a. What is the new correlation? (relevant section)
b. What is the technical term for the finding that this correlation is smaller than the correlation for the full dataset? (relevant section)

The following question is from the Animal Research (AR) case study.

Q23

(AR#11) What is the overall correlation between the belief that animal research is wrong and belief that animal research is necessary? (relevant section)

The following question is from the ADHD Treatment (AT) case study.

Q24

(AT#8) What is the correlation between the participants' correct number of responses after taking the placebo and their correct number of responses after taking 0.60 mg/kg of MPH? (relevant section)

Select Answers

S4

a. 0.58

S8

a. $-4, 2$

S9

a. 109

S17

$r = 0.72$

S19

$r = -0.30$

(larger numbers for the "romint" variable mean less embarrassment)

S21

$r = 0.70$

CHAPTER OVERVIEW

10: Regression

Statisticians are often called upon to develop methods to predict one variable from other variables. For example, one might want to predict college grade point average from high school grade point average. Or, one might want to predict income from the number of years of education.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

10.1: Introduction to Linear Regression

Learning Objectives

- Identify errors of prediction in a scatter plot with a regression line

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the **criterion variable** and is referred to as Y. The variable we are basing our predictions on is called the **predictor variable** and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

The example data in Table $10.1.1$ are plotted in Figure $10.1.1$. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

Table 10.1.1: Example data

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

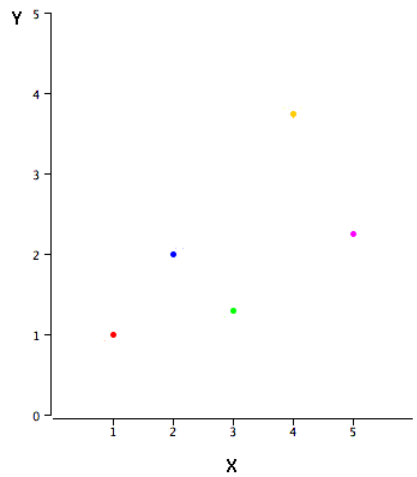


Figure 10.1.1: A scatter plot of the example data

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. The black diagonal line in Figure $10.1.2$ is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

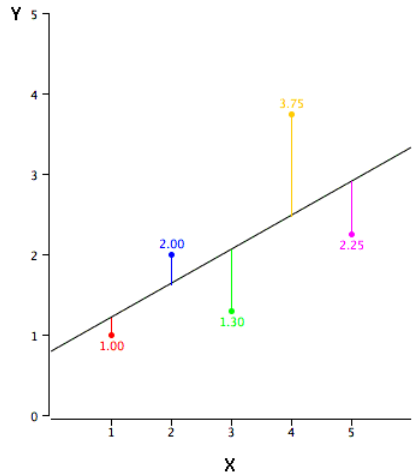


Figure 10.1.2: A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction

The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table $10.1.2$ shows the predicted values (Y') and the errors of prediction ($Y - Y'$). For example, the first point has a Y of 1.00 and a predicted Y (called Y') of 1.21. Therefore, its error of prediction is -0.21.

Table 10.1.2: Example data

X	Y	Y'	Y-Y'	$(Y-Y')^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

You may have noticed that we did not specify what is meant by "best-fitting line." By far, the most commonly-used criterion for the best-fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was used to find the line in Figure $10.1.2$ The last column in Table $10.1.2$ shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table $10.1.2$ is lower than it would be for any other regression line.

The formula for a regression line is

$$Y' = bX + A \tag{10.1.1}$$

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept. The equation for the line in Figure $10.1.2$ is

$$Y' = 0.425X + 0.785 \tag{10.1.2}$$

For $X = 1$,

$$Y' = (0.425)(1) + 0.785 = 1.21 \tag{10.1.3}$$

For $X = 2$,

$$Y' = (0.425)(2) + 0.785 = 1.64 \tag{10.1.4}$$

Computing the Regression Line

In the age of computers, the regression line is typically computed with statistical software. However, the calculations are relatively easy, and are given here for anyone who is interested. The calculations are based on the statistics shown in Table $10.1.3$ M_X is the mean of X, M_Y is the mean of Y, s_X is the standard deviation of X, s_Y is the standard deviation of Y, and r is the correlation between X and Y.

Formula for standard deviation

Formula for correlation

Table 10.1.3: Statistics for computing the regression line

M_X	M_Y	s_X	s_Y	r
3	2.06	1.581	1.072	0.627

The slope (b) can be calculated as follows:

$$b = r\frac{s_Y}{s_X} \tag{10.1.5}$$

and the intercept (A) can be calculated as

$$A = M_Y - bM_X \tag{10.1.6}$$

For these data,

$$b = \frac{(0.627)(1.072)}{1.581} = 0.425 \tag{10.1.7}$$

$$A = 2.06 - (0.425)(3) = 0.785 \tag{10.1.8}$$

Note that the calculations have all been shown in terms of sample statistics rather than population parameters. The formulas are the same; simply use the parameter values for means, standard deviations, and the correlation.

Standardized Variables

The regression equation is simpler if variables are standardized so that their means are equal to 0 and standard deviations are equal to 1, for then $b = r$ and $A = 0$. This makes the regression line:

$$Z_{Y'} = (r)(Z_X) \tag{10.1.9}$$

where $Z_{Y'}$ is the predicted standard score for Y, r is the correlation, and Z_X is the standardized score for X. Note that the slope of the regression equation for standardized variables is r.

A Real Example

The case study "SAT and College GPA" contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university GPA if we knew his or her high school GPA.

Figure $10.1.3$ shows a scatter plot of University GPA as a function of High School GPA. You can see from the figure that there is a strong positive relationship. The correlation is 0.78. The regression equation is:

$$\text{University GPA'} = (0.675)(\text{High School GPA}) + 1.097 \tag{10.1.10}$$

Therefore, a student with a high school GPA of 3 would be predicted to have a university GPA of

$$\text{University GPA'} = (0.675)(3) + 1.097 = 3.12 \tag{10.1.11}$$

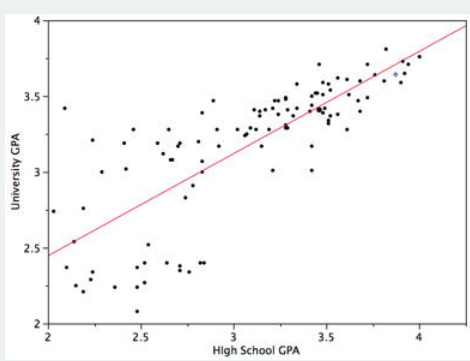


Figure 10.1.3: University GPA as a function of High School GPA

Assumptions

It may surprise you, but the calculations shown in this section are assumption-free. Of course, if the relationship between X and Y were not linear, a different shaped function could fit the data better. Inferential statistics in regression are based on several assumptions, and these assumptions are presented in a later section of this chapter.

10.2: Linear Fit Demo

Learning Objectives

- State the relationship between MSE and fit

Instructions

This demonstration allows you to explore fitting data with linear functions. When the demonstration begins, five points are plotted in the graph. The X axis ranges from 1 to 5 and the Y axis ranges from 0 to 5. The five points are plotted in different colors; next to each point is the Y value of that point. For example, the red point has the value 1.00 next to it. A vertical black line is drawn with the Y value of 3.0; this line consists of the predicted values for Y. (It is clear that this line does not contain the best predictions.) This line is called the "regression line." The equation for the regression line is $Y' = 0X + 3$ where Y' is the predicted value for Y. Since the slope is 0, the same prediction of 3 is made for all values of X.

The error of prediction for each point is represented by a vertical line between the point and the regression line. For the point with a value of 1 on the X axis, the line goes from the point $(1, 1)$ to the point $(1, 3)$ on the line. The length of the vertical line is 2. This means the error of prediction is 2 and the squared error of prediction is $2 \times 2 = 4$. This error is depicted graphically by the graph on the right. The height of the red square is the error of prediction and the area of the red square is the squared error of prediction.

The errors associated with the other points are plotted in a similar way. Therefore the height of the stacked squares is the sum of the errors of prediction (the lengths of the lines are used, so all errors are positive) and the area of all squares is the total sum of squared errors.

This demonstration allows you to change the regression line and examine the effects on the errors of prediction. If you click and drag an end of the line, the slope of the line will change. If you click and drag the middle of the line, the intercept will change. As the line changes, the errors and squared errors are updated automatically.

You can also change the values of the points by clicking on them and dragging. You can only change the Y values.

You can get a good feeling for the regression line and error by changing the points and the slope and intercept of the line and observing the results.

To see the line that minimizes the squared errors of prediction click the "OK" button.

1. Notice that the total deviation is 6. Compute this by summing the 5 separate absolute deviations ($|1-3| + |2-3|$ etc.)
2. The total area is 10. Compute this by summing the squared deviations.
3. Click middle of the black line and drag it so that it is at 4. Has the error increased or decreased?
4. Click the left end of the line and drag it until the intercept is about 1. How has this affected the error?
5. Drag the line further so it has an intercept of about 0. Then drag the upper portion of the line so that it goes through the point at 5.0. Within rounding error, the error should be 0 and the equation for the line should be $Y' = 1X + 0$.
6. Click on the green point at 3.0 and drag it down so that its value is about 2.0. Notice the error now is all based on this one point.
7. Adjust the line to see if you can make the absolute error any smaller.
8. Adjust the line to see of you can make the squared error (area) as small as you can and note equation of the line and the size of the area. Press the OK button to see the smallest possible area. How does the line compare with the one you chose?
9. Move all the points around and then try again to fine the line that makes the area smallest. Compare the area to the line that gives the smallest area.
10. After you have found the line that gives the smallest area (smallest squared error) see if you can change the line so that the absolute error is lower than for this line that gives the smallest area.

Illustrated Instructions

Video Demo

The demonstration starts by dragging each of the 5 points to different locations on the Y axis. Notice how these changes influence the total deviation and area. The video continues by repositioning the regression line by dragging either end as well as the middle. Finally the regression line that minimizes the squared errors is found by clicking the "OK" button.

10.3: Partitioning Sums of Squares

Learning Objectives

- Compute the sum of squares Y
- Convert raw scores to deviation scores
- Compute predicted scores from a regression equation
- Partition sum of squares Y into sum of squares predicted and sum of squares error
- Define r^2 in terms of sum of squares explained and sum of squares Y

One useful aspect of regression is that it can divide the variation in Y into two parts: the variation of the predicted scores and the variation of the errors of prediction. The variation of Y is called the sum of squares Y and is defined as the sum of the squared deviations of Y from the mean of Y. In the population, the formula is

$$SSY = \sum (Y - \mu_Y)^2 \tag{10.3.1}$$

where SSY is the sum of squares Y, Y is an individual value of Y, and μ_y is the mean of Y. A simple example is given in Table $10.3.1$. The mean of Y is 2.06 and SSY is the sum of the values in the third column and is equal to 4.597.

Table 10.3.1: *Example of SSY*

Y	Y-M_y	(Y-M_y)^2
1.00	-1.06	1.1236
2.00	-0.06	0.0036
1.30	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361

When computed in a sample, you should use the sample mean, M, in place of the population mean:

$$SSY = \sum (Y - M_Y)^2 \tag{10.3.2}$$

It is sometimes convenient to use formulas that use deviation scores rather than raw scores. Deviation scores are simply deviations from the mean. By convention, small letters rather than capitals are used for deviation scores. Therefore the score y indicates the difference between Y and the mean of Y. Table $10.3.2$ shows the use of this notation. The numbers are the same as in Table $10.3.1$.

Table 10.3.2: *Example of SSY using Deviation Scores*

Y	**y**	**y^2**
1.00	-1.06	1.1236
2.00	-0.06	0.0036
1.30	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361

The data in Table $10.3.3$ are reproduced from the introductory section. The column X has the values of the predictor variable and the column Y has the criterion variable. The third column, y, contains the differences between the column Y and the mean of Y.

Table 10.3.4: *Example data (The last row contains column sums)*

X	**Y**	**y**	**y^2**	**Y'**	**y'**	**y'^{2}**	**Y-Y'**	**(Y-Y')^2**

X	Y	y	y^2	Y'	y'	y'^2	Y-Y'	$(Y-Y')^2$
1.00	1.00	-1.06	1.1236	1.210	-0.850	0.7225	-0.210	0.044
2.00	2.00	-0.06	0.0036	1.635	-0.425	0.1806	0.365	0.133
3.00	1.30	-0.76	0.5776	2.060	0.000	0.0000	-0.760	0.578
4.00	3.75	1.69	2.8561	2.485	0.425	0.1806	1.265	1.600
5.00	2.25	0.19	0.0361	2.910	0.850	0.7225	-0.660	0.436
Sums								
15.00	10.30	0.00	4.597	10.300	0.000	1.806	0.000	2.791

The fourth column, y^2, is simply the square of the y column. The column Y' contains the predicted values of Y. In the introductory section, it was shown that the equation for the regression line for these data is

$$Y' = 0.425X + 0.785. \tag{10.3.3}$$

The values of Y' were computed according to this equation. The column y' contains deviations of Y' from the mean of Y' and y'^2 is the square of this column. The next-to-last column, $Y - Y'$, contains the actual scores (Y) minus the predicted scores (Y'). The last column contains the squares of these errors of prediction.

We are now in a position to see how the SSY is partitioned. Recall that SSY is the sum of the squared deviations from the mean. It is therefore the sum of the y^2 column and is equal to 4.597. SSY can be partitioned into two parts: the sum of squares predicted (SSY') and the sum of squares error (SSE). The sum of squares predicted is the sum of the squared deviations of the predicted scores from the mean predicted score. In other words, it is the sum of the y'^2 column and is equal to 1.806. The sum of squares error is the sum of the squared errors of prediction. It is therefore the sum of the $(Y - Y')^2$ column and is equal to 2.791. This can be summed up as:

$$SSY = SSY' + SSE \tag{10.3.4}$$

$$4.597 = 1.806 + 2.791 \tag{10.3.5}$$

There are several other notable features about Table $10.3.3$. First, notice that the sum of y and the sum of y' are both zero. This will always be the case because these variables were created by subtracting their respective means from each value. Also, notice that the mean of $Y - Y'$ is 0. This indicates that although some Y values are higher than their respective predicted Y values and some are lower, the average difference is zero.

The SSY is the total variation, the SSY' is the variation explained, and the SSE is the variation unexplained. Therefore, the proportion of variation explained can be computed as:

$$\text{Proportion explained} = \frac{SSY'}{SSY} \tag{10.3.6}$$

Similarly, the proportion not explained is:

$$\text{Proportion not explained} = \frac{SSE}{SSY} \tag{10.3.7}$$

There is an important relationship between the proportion of variation explained and Pearson's correlation: r^2 is the proportion of variation explained. Therefore, if $r = 1$, then, naturally, the proportion of variation explained is 1; if $r = 0$, then the proportion explained is 0. One last example: for $r = 0.4$, the proportion of variation explained is 0.16.

Since the variance is computed by dividing the variation by N (for a population) or $N - 1$ (for a sample), the relationships spelled out above in terms of variation also hold for variance. For example,

$$\sigma^2_{total} = \sigma^2_{Y'} + \sigma^2_e \tag{10.3.8}$$

where the first term is the variance total, the second term is the variance of Y', and the last term is the variance of the errors of prediction ($Y - Y'$). Similarly, r^2 is the proportion of variance explained as well as the proportion of variation explained.

Summary Table

It is often convenient to summarize the partitioning of the data in a table. The degrees of freedom column (df) shows the degrees of freedom for each source of variation. The degrees of freedom for the sum of squares explained is equal to the number of predictor variables. This will always be 1 in simple regression. The error degrees of freedom is equal to the total number of observations minus 2. In this example, it is $5 - 2 = 3$. The total degrees of freedom is the total number of observations minus 1.

Table 10.3.4: *Summary Table for Example Data*

Source	Sum of Squares	df	Mean Square
Explained	1.806	1	1.806
Error	2.791	3	0.930
Total	4.597	4	

Contributor

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

10.4: Standard Error of the Estimate

Learning Objectives

- Make judgments about the size of the standard error of the estimate from a scatter plot
- Compute the standard error of the estimate based on errors of prediction
- Compute the standard error using Pearson's correlation
- Estimate the standard error of the estimate based on a sample

Figure $10.4.1$ shows two regression examples. You can see that in Graph A, the points are closer to the line than they are in Graph B. Therefore, the predictions in Graph A are more accurate than in Graph B.

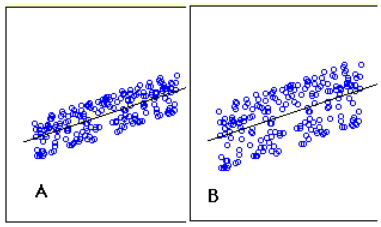


Figure 10.4.1: Regressions differing in accuracy of prediction

The standard error of the estimate is a measure of the accuracy of predictions. Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the sum of squares error). The standard error of the estimate is closely related to this quantity and is defined below:

$$\sigma_{est} = \sqrt{\frac{\sum(Y-Y')^2}{N}} \tag{10.4.1}$$

where σ_{est} is the standard error of the estimate, Y is an actual score, Y' is a predicted score, and N is the number of pairs of scores. The numerator is the sum of squared differences between the actual scores and the predicted scores.

Note the similarity of the formula for σ_{est} to the formula for σ. It turns out that σest is the standard deviation of the errors of prediction (each $Y-Y'$ is an error of prediction).

Assume the data in Table $10.4.1$ are the data from a population of five X, Y pairs.

Table 10.4.1: Example data

	X	Y	Y'	Y-Y'	$(Y-Y')^2$
	1.00	1.00	1.210	-0.210	0.044
	2.00	2.00	1.635	0.365	0.133
	3.00	1.30	2.060	-0.760	0.578
	4.00	3.75	2.485	1.265	1.600
	5.00	2.25	2.910	-0.660	0.436
Sum	15.00	10.30	10.30	0.000	2.791

The last column shows that the sum of the squared errors of prediction is 2.791. Therefore, the standard error of the estimate is

$$\sigma_{est} = \sqrt{\frac{2.791}{5}} = 0.747 \tag{10.4.2}$$

There is a version of the formula for the standard error in terms of Pearson's correlation:

$$\sigma_{est} = \sqrt{\frac{(1-\rho)^2 SSY}{N}} \tag{10.4.3}$$

where ρ is the population value of Pearson's correlation and SSY is

$$SSY = \sum (Y - \mu_Y)^2 \tag{10.4.4}$$

For the data in Table 10.4.1, $\mu_Y = 2.06$, $SSY = 4.597$ and $\rho = 0.6268$. Therefore,

$$\sigma_{est} = \sqrt{\frac{(1 - 0.6268^2)(4.597)}{5}} = \sqrt{\frac{2.791}{5}} = 0.747 \tag{10.4.5}$$

which is the same value computed previously.

Similar formulas are used when the standard error of the estimate is computed from a sample rather than a population. The only difference is that the denominator is $N - 2$ rather than N. The reason $N - 2$ is used rather than $N - 1$ is that two parameters (the slope and the intercept) were estimated in order to estimate the sum of squares. Formulas for a sample comparable to the ones for a population are shown below.

$$s_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N - 2}} \tag{10.4.6}$$

$$s_{est} = \sqrt{\frac{2.791}{3}} = 0.964 \tag{10.4.7}$$

$$s_{est} = \sqrt{\frac{(1 - r)^2 SSY}{N - 2}} \tag{10.4.8}$$

10.5: Inferential Statistics for b and r

Learning Objectives

- State the assumptions that inferential statistics in regression are based upon
- Identify heteroscedasticity in a scatter plot
- Compute the standard error of a slope
- Test a slope for significance
- Construct a confidence interval on a slope
- Test a correlation for significance

This section shows how to conduct significance tests and compute confidence intervals for the regression slope and Pearson's correlation. As you will see, if the regression slope is significantly different from zero, then the correlation coefficient is also significantly different from zero.

Assumptions

Although no assumptions were needed to determine the best-fitting straight line, assumptions are made in the calculation of inferential statistics. Naturally, these assumptions refer to the population, not the sample.

1. **Linearity**: The relationship between the two variables is linear.
2. **Homoscedasticity**: The variance around the regression line is the same for all values of X. A clear violation of this assumption is shown in Figure $10.5.1$. Notice that the predictions for students with high high-school GPAs are very good, whereas the predictions for students with low high-school GPAs are not very good. In other words, the points for students with high high-school GPAs are close to the regression line, whereas the points for low high-school GPA students are not.

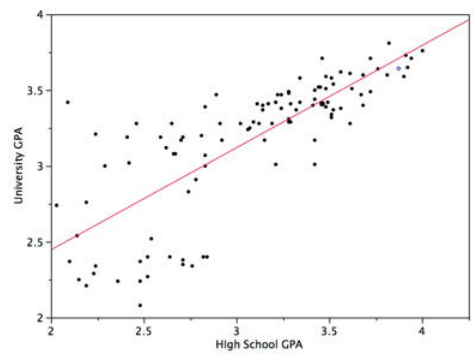


Figure 10.5.1: University GPA as a function of High School GPA

3. The errors of prediction are distributed normally. This means that the deviations from the regression line are normally distributed. It does not mean that X or Y is normally distributed.

Significance Test for the Slope (b)

Recall the general formula for a t test:

$$t = \frac{\text{statistic-hypothesized value}}{\text{estimated standard error of the statistic}} \tag{10.5.1}$$

As applied here, the statistic is the sample value of the slope (b) and the hypothesized value is 0. The number of degrees of freedom for this test is:

$$df = N - 2 \tag{10.5.2}$$

where N is the number of pairs of scores.

The estimated standard error of b is computed using the following formula:

$$s_b = \frac{s_{est}}{\sqrt{SSX}} \tag{10.5.3}$$

where s_b is the estimated standard error of b, s_{est} is the standard error of the estimate, and SSX is the sum of squared deviations of X from the mean of X. SSX is calculated as

$$SSX = \sum (X - M_X)^2 \tag{10.5.4}$$

where M_x is the mean of X. As shown previously, the standard error of the estimate can be calculated as

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}} \tag{10.5.5}$$

These formulas are illustrated with the data shown in Table $10.5.1$. These data are reproduced from the introductory section. The column X has the values of the predictor variable and the column Y has the values of the criterion variable. The third column, x, contains the differences between the values of column X and the mean of X. The fourth column, x_2, is the square of the x column. The fifth column, y, contains the differences between the values of column Y and the mean of Y. The last column, y_2, is simply square of the y column.

Table $10.5.1$: Example data

	X	Y	x	x2	y	y2
	1.00	1.00	-2.00	4	-1.06	1.1236
	2.00	2.00	-1.00	1	-0.06	0.0036
	3.00	1.30	0.00	0	-0.76	0.5776
	4.00	3.75	1.00	1	1.69	2.8561
	5.00	2.25	2.00	4	0.19	0.0361
Sum	15.00	10.30	0.00	10.00	0.00	4.5970

The computation of the standard error of the estimate (s_{est}) for these data is shown in the section on the standard error of the estimate. It is equal to 0.964.

$$s_{est} = 0.964 \tag{10.5.6}$$

SSX is the sum of squared deviations from the mean of X. It is, therefore, equal to the sum of the x^2 column and is equal to 10.

$$SSX = 10.00 \tag{10.5.7}$$

We now have all the information to compute the standard error of b:

$$s_b = \frac{0.964}{\sqrt{10}} = 0.305 \tag{10.5.8}$$

As shown previously, the slope (b) is 0.425. Therefore,

$$t = \frac{0.425}{0.305} = 1.39 \tag{10.5.9}$$

$$df = N - 2 = 5 - 2 = 3 \tag{10.5.10}$$

The p value for a two-tailed t test is 0.26. Therefore, the slope is not significantly different from 0.

Confidence Interval for the Slope

The method for computing a confidence interval for the population slope is very similar to methods for computing other confidence intervals. For the 95% confidence interval, the formula is:

$$\text{lower limit}: b - (t_{0.95})(s_b) \tag{10.5.11}$$

$$\text{upper limit}: b + (t_{0.95})(s_b) \tag{10.5.12}$$

where $t_{0.95}$ is the value of t to use for the 95% confidence interval.

The values of t to be used in a confidence interval can be looked up in a table of the t distribution. A small version of such a table is shown in Table $10.5.2$ The first column, df, stands for degrees of freedom.

Table $10.5.2$: Abbreviated t table

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

You can also use the "inverse t distribution" calculator to find the t values to use in a confidence interval.

Applying these formulas to the example data,

$$\text{lower limit}: 0.425 - (3.182)(0.305) = -0.55 \tag{10.5.13}$$

$$\text{upper limit}: 0.425 + (3.182)(0.305) = 1.40 \tag{10.5.14}$$

Significance Test for the Correlation

The formula for a significance test of Pearson's correlation is shown below:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{10.5.15}$$

where N is the number of pairs of scores. For the example data,

$$t = \frac{0.627\sqrt{5-2}}{\sqrt{1-0.627^2}} = 1.39 \tag{10.5.16}$$

Notice that this is the same t value obtained in the t test of b. As in that test, the degrees of freedom is $N-2 = 5-2 = 3$.

10.6: Influential Observations

Learning Objectives

- Describe what makes a point influential
- Define "distance"

It is possible for a single observation to have a great influence on the results of a regression analysis. It is therefore important to be alert to the possibility of influential observations and to take them into consideration when interpreting the results.

Influence

The influence of an observation can be thought of in terms of how much the predicted scores for other observations would differ if the observation in question were not included. **Cook's D** is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question. If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

A common rule of thumb is that an observation with a value of **Cook's D** over 1.0 has too much influence. As with all rules of thumb, this rule should be applied judiciously and not thoughtlessly.

An observation's influence is a function of two factors:

1. How much the observation's value on the predictor variable differs from the mean of the predictor variable and
2. The difference between the predicted score for the observation and its actual score.

The former factor is called the observation's leverage. The latter factor is called the observation's distance.

Calculation of Cook's D (Optional)

The first step in calculating the value of **Cook's D** for an observation is to predict all the scores in the data once using a regression equation based on all the observations and once using all the observations except the observation in question. The second step is to compute the sum of the squared differences between these two sets of predictions. The final step is to divide this result by 2 times the MSE (see the section on partitioning the variance).

Leverage

The leverage of an observation is based on how much the observation's value on the predictor variable differs from the mean of the predictor variable. The greater an observation's leverage, the more potential it has to be an influential observation. For example, an observation with a value equal to the mean on the predictor variable has no influence on the slope of the regression line regardless of its value on the criterion variable. On the other hand, an observation that is extreme on the predictor variable has the potential to affect the slope greatly.

Calculation of Leverage (h)

The first step is to standardize the predictor variable so that it has a mean of 0 and a standard deviation of 1. Then, the leverage (h) is computed by squaring the observation's value on the standardized predictor variable, adding 1, and dividing by the number of observations.

Distance

The distance of an observation is based on the error of prediction for the observation: The greater the error of prediction, the greater the distance. The most commonly used measure of distance is the $studentized\,residual$. The $studentized\,residual$ for an observation is closely related to the error of prediction for that observation divided by the standard deviation of the errors of prediction. However, the predicted score is derived from a regression equation in which the observation in question is not counted. The details of the computation of a $studentized\,residual$ are a bit complex and are beyond the scope of this work.

Even an observation with a large distance will not have that much influence if its leverage is low. It is the combination of an observation's leverage and distance that determines its influence.

Example 10.6.1

Table 10.6.1 shows the leverage, *studentized residual*, and influence for each of the five observations in a small dataset.

Table **10.6.1:** *Example Data*

ID	X	Y	h	R	D
A	1	2	0.39	-1.02	0.40
B	2	3	0.27	-0.56	0.06
C	3	5	0.21	0.89	0.11
D	4	6	0.20	1.22	0.19
E	8	7	0.73	-1.68	8.86

In the above table, h is the leverage, R is the *studentized residual*, and D is Cook's measure of influence.

Observation A has fairly high leverage, a relatively high residual, and moderately high influence.

Observation B has small leverage and a relatively small residual. It has very little influence.

Observation C has small leverage and a relatively high residual. The influence is relatively low.

Observation D has the lowest leverage and the second highest residual. Although its residual is much higher than Observation A , its influence is much less because of its low leverage.

Observation E has by far the largest leverage and the largest residual. This combination of high leverage and high residual makes this observation extremely influential.

Figure 10.6.1 shows the regression line for the whole dataset (blue) and the regression line if the observation in question is not included (red) for all observations. The observation in question is circled. Naturally, the regression line for the whole dataset is the same in all panels. The residual is calculated relative to the line for which the observation in question is not included in the analysis. The most influential observation is Observation E for which the two regression lines are very different. This indicates the influence of this observation.

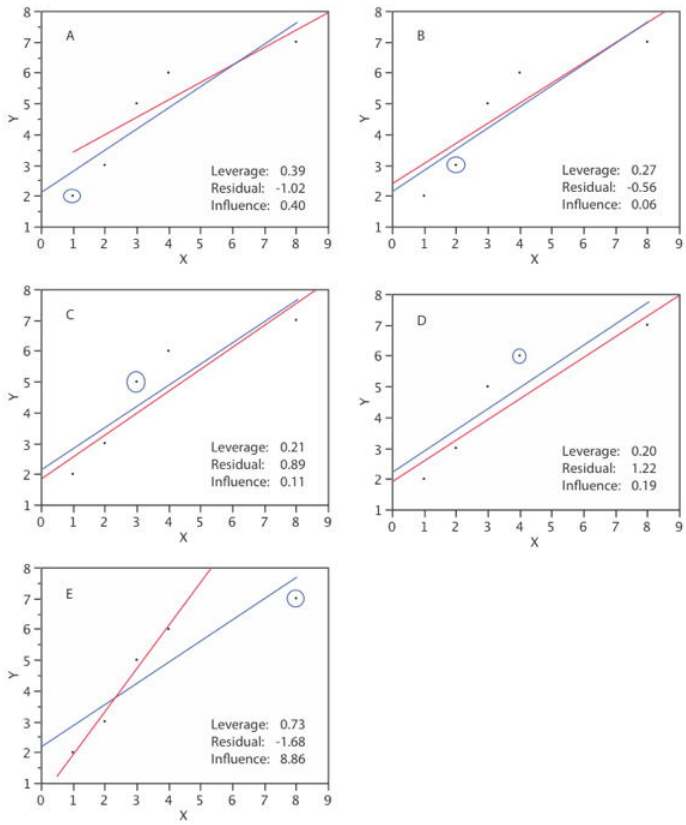


Figure 10.6.1: Illustration of leverage, residual, and influence. The circled points are not included in the calculation of the red regression line. All points are included in the calculation of the blue regression line.

10.7: Regression Toward the Mean

Learning Objectives

- To study Regression toward the mean

Regression toward the mean involves outcomes that are at least partly due to chance. We begin with an example of a task that is entirely chance: Imagine an experiment in which a group of 25 people each predicted the outcomes of flips of a fair coin. For each subject in the experiment, a coin is flipped 12 times and the subject predicts the outcome of each flip. Figure $10.7.1$ shows the results of a simulation of this "experiment." Although most subjects were correct from 5 to 8 times out of 12, one simulated subject was correct 10 times. Clearly, this subject was very lucky and probably would not do as well if he or she performed the task a second time. In fact, the best prediction of the number of times this subject would be correct on the retest is 6, since the probability of being correct on a given trial is 0.5 and there are 12 trials.

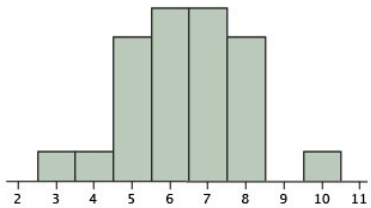


Figure 10.7.1: Histogram of results of a simulated experiment

More technically, the best prediction for the subject's result on the retest is the mean of the binomial distribution with $N = 12$ and $p = 0.50$. This distribution is shown in Figure $10.7.2$ and has a mean of 6.

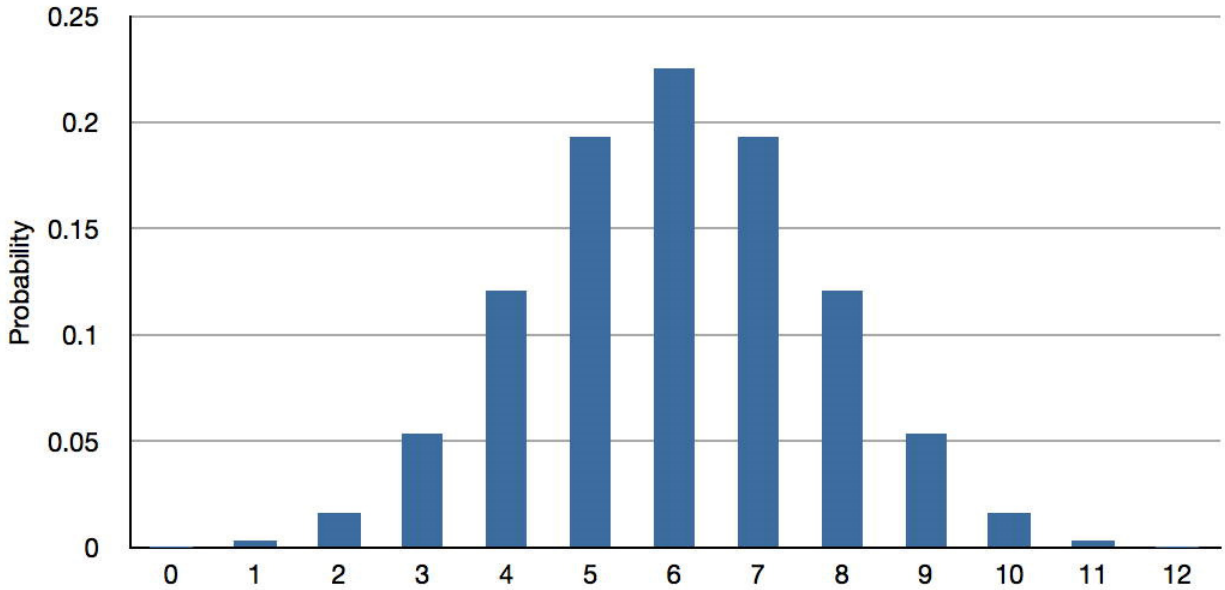


Figure 10.7.2: Binomial Distribution for $N = 12$ and $p = 0.50$

The point here is that no matter how many coin flips a subject predicted correctly, the best prediction of their score on a retest is 6.

Now we consider a test we will call "**Test A**" that is partly chance and partly skill: Instead of predicting the outcomes of 12 coin flips, each subject predicts the outcomes of 6 coin flips and answers 6 true/false questions about world history. Assume that the mean score on the 6 history questions is 4. A subject's score on **Test A** has a large chance component but also depends on history knowledge. If a subject scored very high on this test (such as a score of $10/12$), it is likely that they did well on both the history questions and the coin flips. If this subject is then given a second test (**Test B**) that also included coin predictions and history questions, their knowledge of history would be helpful and they would again be expected to score above the mean. However, since

their high performance on the coin portion of **Test A** would not be predictive of their coin performance on **Test B**, they would not be expected to fare as well on **Test B** as on **Test A**. Therefore, the best prediction of their score on **Test B** would be somewhere between their score on **Test A** and the mean of **Test B**. This tendency of subjects with high values on a measure that includes chance and skill to score closer to the mean on a retest is called "*regression toward the mean*."

The essence of the regression-toward-the-mean phenomenon is that people with high scores tend to be above average in skill and in luck and that only the skill portion is relevant to future performance. Similarly, people with low scores tend to be below average in skill and luck, and their bad luck is not relevant to future performance. This does not mean that all people who score high have above average luck. However, on average they do.

Almost every measure of behavior has a chance and a skill component to it. Take a student's grade on a final exam as an example. Certainly, the student's knowledge of the subject will be a major determinant of his or her grade. However, there are aspects of performance that are due to chance. The exam cannot cover everything in the course and therefore must represent a subset of the material. Maybe the student was lucky in that the one aspect of the course the student did not understand well was not well-represented on the test. Or, maybe the student was not sure which of two approaches to a problem would be better but, more or less by chance, chose the right one. Other chance elements come into play as well. Perhaps the student was awakened early in the morning by a random phone call, resulting in fatigue and lower performance. And, of course, guessing on multiple choice questions is another source of randomness in test scores.

There will be regression toward the mean in a test-retest situation whenever there is less than a perfect ($r = 1$) relationship between the test and the retest. This follows from the formula for a regression line with standardized variables shown below:

$$Z_{Y'} = (r)(Z_X) \tag{10.7.1}$$

From this equation it is clear that if the absolute value of r is less than 1, then the predicted value of Z_Y will be closer to 0, the mean for standardized scores, than is Z_X. Also, note that if the correlation between X and Y is 0, as it would be for a task that is all luck, the predicted standard score for Y is its mean, 0, regardless of the score on X.

Figure $10.7.3$ shows a scatter plot with the regression line predicting the standardized Verbal SAT from the standardized Math SAT. Note that the slope of the line is equal to the correlation of 0.835 between these variables.

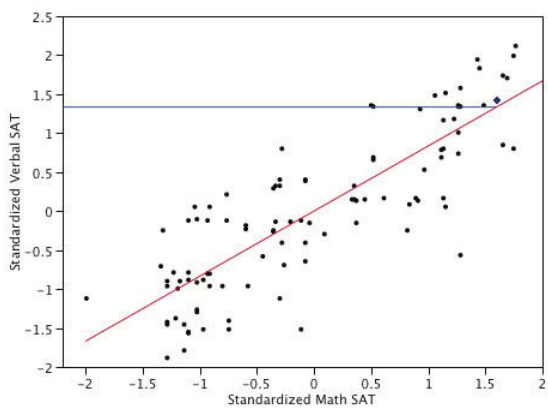


Figure 10.7.3: Prediction of Standardized Verbal SAT from Standardized Math SAT

The point represented by a blue diamond has a value of 1.6 on the standardized Math SAT. This means that this student scored 1.6 standard deviations above the mean on Math SAT. The predicted score is $(r)(1.6) = (0.835)(1.6) = 1.34$. The horizontal line on the graph shows the value of the predicted score. The key point is that although this student scored 1.6 standard deviations above the mean on Math SAT, he or she is only predicted to score 1.34 standard deviations above the mean on Verbal SAT. Thus, the prediction is that the Verbal SAT score will be closer to the mean of 0 than is the Math SAT score. Similarly, a student scoring far below the mean on Math SAT will be predicted to score higher on Verbal SAT.

Regression toward the mean occurs in any situation in which observations are selected on the basis of performance on a task that has a random component. If you choose people on the basis of their performance on such a task, you will be choosing people partly on the basis of their skill and partly on the basis of their luck on the task. Since their luck cannot be expected to be maintained from trial to trial, the best prediction of a person's performance on a second trial will be somewhere between their performance on the first trial and the mean performance on the first trial. The degree to which the score is expected to "regress toward the mean" in this

manner depends on the relative contributions of chance and skill to the task: the greater the role of chance, the more the regression toward the mean.

Errors Resulting From Failure to Understand Regression Toward the Mean

Failure to appreciate regression toward the mean is common and often leads to incorrect interpretations and conclusions. One of the best examples is provided by Nobel Laureate Daniel Kahneman in his autobiography. Dr. Kahneman was attempting to teach flight instructors that praise is more effective than punishment. He was challenged by one of the instructors who relayed that in his experience praising a cadet for executing a clean maneuver is typically followed by a lesser performance, whereas screaming at a cadet for bad execution is typically followed by improved performance. This, of course, is exactly what would be expected based on regression toward the mean. A pilot's performance, although based on considerable skill, will vary randomly from maneuver to maneuver. When a pilot executes an extremely clean maneuver, it is likely that he or she had a bit of luck in their favor in addition to their considerable skill. After the praise but not because of it, the luck component will probably disappear and the performance will be lower. Similarly, a poor performance is likely to be partly due to bad luck. After the criticism but not because of it, the next performance will likely be better. To drive this point home, Kahneman had each instructor perform a task in which a coin was tossed at a target twice. He demonstrated that the performance of those who had done the best the first time deteriorated, whereas the performance of those who had done the worst improved.

Regression toward the mean is frequently present in sports performance. A good example is provided by Schall and Smith (2000), who analyzed many aspects of baseball statistics including the batting averages of players in 1998. They chose the 10 players with the highest batting averages (BAs) in 1998 and checked to see how well they did in 1999. According to what would be expected based on regression toward the mean, these players should, on average, have lower batting averages in 1999 than they did in 1998. As can be seen in Table $10.7.1$, $7/10$ of the players had lower batting averages in 1999 than they did in 1998. Moreover, those who had higher averages in 1999 were only slightly higher, whereas those who were lower were much lower. The average decrease from 1998 to 1999 was 33 points. Even so, most of these players had excellent batting averages in 1999 indicating that skill was an important component of their 1998 averages.

Table $10.7.1$: How the Ten Players with the Highest BAs in 1998 did in 1999

1998	1999	Difference
363	379	16
354	298	-56
339	342	3
337	281	-56
336	249	-87
331	298	-33
328	297	-31
328	303	-25
327	257	-70
327	332	5

Figure $10.7.4$ shows the batting averages of the two years. The decline from 1998 to 1999 is clear. Note that although the mean decreased from 1998, some players increased their batting averages. This illustrates that regression toward the mean does not occur for every individual. Although the predicted scores for every individual will be lower, some of the predictions will be wrong.

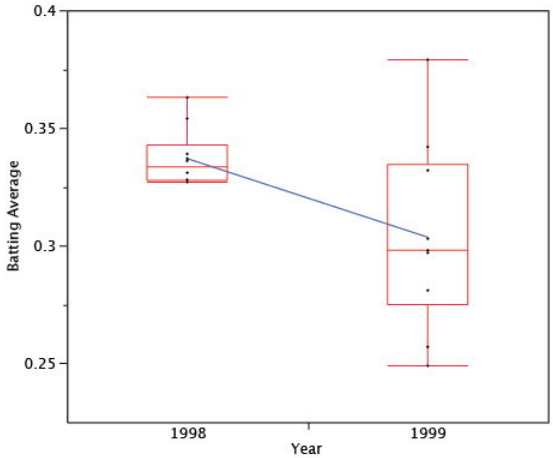


Figure 10.7.4: Quantile plots of the batting averages. The line connects the means of the plots.

Regression toward the mean plays a role in the so-called "Sophomore Slump," a good example of which is that a player who wins "rookie of the year" typically does less well in his second season. A related phenomenon is called the Sports Illustrated Cover Jinx.

An experiment without a control group can confound regression effects with real effects. For example, consider a hypothetical experiment to evaluate a reading-improvement program. All first graders in a school district were given a reading achievement test and the 50 lowest-scoring readers were enrolled in the program. The students were retested following the program and the mean improvement was large. Does this necessarily mean the program was effective? No, it could be that the initial poor performance of the students was due, in part, to bad luck. Their luck would be expected to improve in the retest, which would increase their scores with or without the treatment program.

For a real example, consider an experiment that sought to determine whether the drug propranolol would increase the SAT scores of students thought to have test anxiety. Propranolol was given to 25 high school students chosen because IQ tests and other academic performance indicated that they had not done as well as expected on the SAT. On a retest taken after receiving propranolol, students improved their SAT scores an average of 120 points. This was a significantly greater increase than the 38 points expected simply on the basis of having taken the test before. The problem with the study is that the method of selecting students likely resulted in a disproportionate number of students who had bad luck when they first took the SAT. Consequently, these students would likely have increased their scores on a retest with or without the propranolol. This is not to say that propranolol had no effect. However, since possible propranolol effects and regression effects were confounded, no firm conclusions should be drawn.

Randomly assigning students to either the propranolol group or a control group would have improved the experimental design. Since the regression effects would then not have been systematically different for the two groups, a significant difference would have provided good evidence for a propranolol effect.

New York Times Article on the Study

Schall, T., & Smith, G. (2000) Do Baseball Players Regress Toward the Mean? The American Statistician, 54, 231-235.

10.8: Introduction to Multiple Regression

Learning Objectives

- State the regression equation
- Define "regression coefficient"
- Define "beta weight"
- Explain what R is and how it is related to r
- Explain why a regression weight is called a "partial slope"
- Explain why the sum of squares explained in a multiple regression model is usually less than the sum of the sums of squares in simple regression
- Define R^2 in terms of proportion explained
- Test R^2 for significance
- Test the difference between a complete and reduced model for significance
- State the assumptions of multiple regression and specify which aspects of the analysis require assumptions

In simple linear regression, a criterion variable is predicted from one predictor variable. In multiple regression, the criterion is predicted by two or more variables. For example, in the SAT case study, you might want to predict a student's university grade point average on the basis of their High-School GPA ($HSGPA$) and their total SAT score (verbal + math). The basic idea is to find a linear combination of $HSGPA$ and SAT that best predicts University GPA ($UGPA$). That is, the problem is to find the values of b_1 and b_2 in the equation shown below that give the best predictions of $UGPA$. As in the case of simple linear regression, we define the best predictions as the predictions that minimize the squared errors of prediction.

$$UGPA' = b_1 HSGPA + b_2 SAT + A \tag{10.8.1}$$

where $UGPA'$ is the predicted value of University GPA and A is a constant. For these data, the best prediction equation is shown below:

$$UGPA' = 0.541 \times HSGPA + 0.008 \times SAT + 0.540 \tag{10.8.2}$$

In other words, to compute the prediction of a student's University GPA, you add up their High-School GPA multiplied by 0.541, their SAT multiplied by 0.008, and 0.540. Table $10.8.1$ shows the data and predictions for the first five students in the dataset.

Table $10.8.1$: Data and Predictions

HSGPA	SAT	UGPA'
3.45	1232	3.38
2.78	1070	2.89
2.52	1086	2.76
3.67	1287	3.55
3.24	1130	3.19

The values of b (b_1 and b_2) are sometimes called "regression coefficients" and sometimes called "regression weights." These two terms are synonymous.

The multiple correlation (R) is equal to the correlation between the predicted scores and the actual scores. In this example, it is the correlation between $UGPA'$ and $UGPA$, which turns out to be 0.79. That is, $R = 0.79$. Note that R will never be negative since if there are negative correlations between the predictor variables and the criterion, the regression weights will be negative so that the correlation between the predicted and actual scores will be positive.

Interpretation of Regression Coefficients

A regression coefficient in multiple regression is the slope of the linear relationship between the criterion variable and the part of a predictor variable that is independent of all other predictor variables. In this example, the regression coefficient for $HSGPA$ can be computed by first predicting $HSGPA$ from SAT and saving the errors of prediction (the differences between $HSGPA$ and $HSGPA'$). These errors of prediction are called "residuals" since they are what is left over in $HSGPA$ after the predictions from SAT are subtracted, and represent the part of $HSGPA$ that is independent of SAT. These residuals are referred to as $HSGPA.SAT$, which means they are the residuals in $HSGPA$ after having been predicted by SAT. The correlation between $HSGPA.SAT$ and SAT is necessarily 0.

The final step in computing the regression coefficient is to find the slope of the relationship between these residuals and $UGPA$. This slope is the regression coefficient for $HSGPA$. The following equation is used to predict $HSGPA$ from SAT:

$$HSGPA' = -1.314 + 0.0036 \times SAT \tag{10.8.3}$$

The residuals are then computed as:

$$HSGPA - HSGPA' \tag{10.8.4}$$

The linear regression equation for the prediction of $UGPA$ by the residuals is

$$UGPA' = 0.541 \times HSGPA.SAT + 3.173 \tag{10.8.5}$$

Notice that the slope (0.541) is the same value given previously for b_1 in the multiple regression equation.

This means that the regression coefficient for $HSGPA$ is the slope of the relationship between the criterion variable and the part of $HSGPA$ that is independent of (uncorrelated with) the other predictor variables. It represents the change in the criterion variable associated with a change of one in the predictor variable when all other predictor variables are held constant. Since the regression coefficient for $HSGPA$ is 0.54, this means that, holding SAT constant, a change of one in $HSGPA$ is associated with a change of 0.54 in $UGPA'$. If two students had the same SAT and differed in $HSGPA$ by 2, then you would predict they would differ in $UGPA$ by $(2)(0.54) = 1.08$. Similarly, if they differed by 0.5, then you would predict they would differ by $(0.50)(0.54) = 0.27$.

The slope of the relationship between the part of a predictor variable independent of other predictor variables and the criterion is its partial slope. Thus the regression coefficient of 0.541 for $HSGPA$ and the regression coefficient of 0.008 for SAT are partial slopes. Each partial slope represents the relationship between the predictor variable and the criterion holding constant all of the other predictor variables.

It is difficult to compare the coefficients for different variables directly because they are measured on different scales. A difference of 1 in $HSGPA$ is a fairly large difference, whereas a difference of 1 on the SAT is negligible. Therefore, it can be advantageous to transform the variables so that they are on the same scale. The most straightforward approach is to standardize the variables so that they each have a standard deviation of 1. A regression weight for standardized variables is called a "beta weight" and is designated by the Greek letter β. For these data, the beta weights are 0.625 and 0.198. These values represent the change in the criterion (in standard deviations) associated with a change of one standard deviation on a predictor [holding constant the value(s) on the other predictor(s)]. Clearly, a change of one standard deviation on $HSGPA$ is associated with a larger difference than a change of one standard deviation of SAT. In practical terms, this means that if you know a student's $HSGPA$, knowing the student's SAT does not aid the prediction of $UGPA$ much. However, if you do not know the student's $HSGPA$, his or her SAT can aid in the prediction since the β weight in the simple regression predicting $UGPA$ from SAT is 0.68. For comparison purposes, the β weight in the simple regression predicting $UGPA$ from $HSGPA$ is 0.78. As is typically the case, the partial slopes are smaller than the slopes in simple regression.

Partitioning the Sums of Squares

Just as in the case of simple linear regression, the sum of squares for the criterion ($UGPA$ in this example) can be partitioned into the sum of squares predicted and the sum of squares error. That is,

$$SSY = SSY' + SSE \tag{10.8.6}$$

which for these data:

$$20.798 = 12.961 + 7.837 \tag{10.8.7}$$

The sum of squares predicted is also referred to as the "sum of squares explained." Again, as in the case of simple regression,

$$\text{Proportion Explained} = SSY'/SSY \tag{10.8.8}$$

In simple regression, the proportion of variance explained is equal to r^2; in multiple regression, the proportion of variance explained is equal to R^2.

In multiple regression, it is often informative to partition the sum of squares explained among the predictor variables. For example, the sum of squares explained for these data is 12.96. How is this value divided between $HSGPA$ and SAT? One approach that, as will be seen, does not work is to predict $UGPA$ in separate simple regressions for $HSGPA$ and SAT. As can be seen in Table $10.8.2$, the sum of squares in these separate simple regressions is 12.64 for $HSGPA$ and 9.75 for SAT. If we add these two sums of squares we get 22.39, a value much larger than the sum of squares explained of 12.96 in the multiple regression analysis. The explanation is that $HSGPA$ and SAT are highly correlated ($r = 0.78$) and therefore much of the variance in $UGPA$ is confounded between $HSGPA$ and SAT. That is, it could be explained by either $HSGPA$ or SAT and is counted twice if the sums of squares for $HSGPA$ and SAT are simply added.

Table $10.8.2$: Sums of Squares for Various Predictors

Predictors	Sum of Squares
HSGPA	12.64
SAT	9.75
HSGPA and SAT	12.96

Table $10.8.3$ shows the partitioning of the sum of squares into the sum of squares uniquely explained by each predictor variable, the sum of squares confounded between the two predictor variables, and the sum of squares error. It is clear from this table that most of the sum of squares explained is confounded between $HSGPA$ and SAT. Note that the sum of squares uniquely explained by a predictor variable is analogous to the partial slope of the variable in that both involve the relationship between the variable and the criterion with the other variable(s) controlled.

Table $10.8.3$: Partitioning the Sum of Squares

Source	Sum of Squares	Proportion
HSGPA (unique)	3.21	0.15
SAT (unique)	0.32	0.02
HSGPA and SAT (Confounded)	9.43	0.45
Error	7.84	0.38
Total	20.80	1.00

The sum of squares uniquely attributable to a variable is computed by comparing two regression models: the complete model and a reduced model. The complete model is the multiple regression with all the predictor variables included ($HSGPA$ and SAT in this example). A reduced model is a model that leaves out one of the predictor variables. The sum of squares uniquely attributable to a variable is the sum of squares for the complete model minus the sum of squares for the reduced model in which the variable of interest is omitted. As shown in Table $10.8.2$, the sum of squares for the complete model ($HSGPA$ and SAT) is 12.96. The sum of squares for the reduced model in which $HSGPA$ is omitted is simply the sum of squares explained using SAT as the predictor variable and is 9.75. Therefore, the sum of squares uniquely attributable to $HSGPA$ is $12.96 - 9.75 = 3.21$. Similarly, the sum of squares uniquely attributable to SAT is $12.96 - 12.64 = 0.32$. The confounded sum of squares in this example is computed by subtracting the sum of squares uniquely attributable to the predictor variables from the sum of squares for the complete model:

$12.96 - 3.21 - 0.32 = 9.43$. The computation of the confounded sums of squares in analysis with more than two predictors is more complex and beyond the scope of this text.

Since the variance is simply the sum of squares divided by the degrees of freedom, it is possible to refer to the proportion of variance explained in the same way as the proportion of the sum of squares explained. It is slightly more common to refer to the proportion of variance explained than the proportion of the sum of squares explained and, therefore, that terminology will be adopted frequently here.

When variables are highly correlated, the variance explained uniquely by the individual variables can be small even though the variance explained by the variables taken together is large. For example, although the proportions of variance explained uniquely by $HSGPA$ and SAT are only 0.15 and 0.02 respectively, together these two variables explain 0.62 of the variance. Therefore, you could easily underestimate the importance of variables if only the variance explained uniquely by each variable is considered. Consequently, it is often useful to consider a set of related variables. For example, assume you were interested in predicting job performance from a large number of variables some of which reflect cognitive ability. It is likely that these measures of cognitive ability would be highly correlated among themselves and therefore no one of them would explain much of the variance independently of the other variables. However, you could avoid this problem by determining the proportion of variance explained by all of the cognitive ability variables considered together as a set. The variance explained by the set would include all the variance explained uniquely by the variables in the set as well as all the variance confounded among variables in the set. It would not include variance confounded with variables outside the set. In short, you would be computing the variance explained by the set of variables that is independent of the variables not in the set.

Inferential Statistics

We begin by presenting the formula for testing the significance of the contribution of a set of variables. We will then show how special cases of this formula can be used to test the significance of R^2 as well as to test the significance of the unique contribution of individual variables.

The first step is to compute two regression analyses:

1. an analysis in which all the predictor variables are included and
2. an analysis in which the variables in the set of variables being tested are excluded.

The former regression model is called the "complete model" and the latter is called the "reduced model." The basic idea is that if the reduced model explains much less than the complete model, then the set of variables excluded from the reduced model is important.

The formula for testing the contribution of a group of variables is:

$$F = \frac{\dfrac{SSQ_C - SSQ_R}{p_C - p_R}}{\dfrac{SSQ_T - SSQ_C}{N - p_C - 1}} = \frac{MS_{explained}}{MS_{error}} \tag{10.8.9}$$

where:

SSQ_C is the sum of squares for the complete model,

SSQ_R is the sum of squares for the reduced model,

p_C is the number of predictors in the complete model,

p_R is the number of predictors in the reduced model,

SSQ_T is the sum of squares total (the sum of squared deviations of the criterion variable from its mean), and

N is the total number of observations

The degrees of freedom for the numerator is $p_C - p_R$ and the degrees of freedom for the denominator is $N - p_C - 1$. If the F is significant, then it can be concluded that the variables excluded in the reduced set contribute to the prediction of the criterion variable independently of the other variables.

This formula can be used to test the significance of R^2 by defining the reduced model as having no predictor variables. In this application, SSQ_R and $p_R = 0$. The formula is then simplified as follows:

$$F = \frac{\frac{SSQ_C}{p_C}}{\frac{SSQ_T - SSQ_C}{N - p_C - 1}} = \frac{MS_{explained}}{MS_{error}} \tag{10.8.10}$$

which for this example becomes:

$$F = \frac{\frac{12.96}{2}}{\frac{20.80 - 12.96}{105 - 2 - 1}} = \frac{6.48}{0.08} = 84.35 \tag{10.8.11}$$

The degrees of freedom are 2 and 102. The F distribution calculator shows that $p < 0.001$.

F Calculator

The reduced model used to test the variance explained uniquely by a single predictor consists of all the variables except the predictor variable in question. For example, the reduced model for a test of the unique contribution of $HSGPA$ contains only the variable SAT. Therefore, the sum of squares for the reduced model is the sum of squares when $UGPA$ is predicted by SAT. This sum of squares is 9.75. The calculations for F are shown below:

$$F = \frac{\frac{12.96 - 9.75}{2 - 1}}{\frac{20.80 - 12.96}{105 - 2 - 1}} = \frac{3.212}{0.077} = 41.80 \tag{10.8.12}$$

The degrees of freedom are 1 and 102. The F distribution calculator shows that $p < 0.001$.

Similarly, the reduced model in the test for the unique contribution of SAT consists of $HSGPA$.

$$F = \frac{\frac{12.96 - 12.64}{2 - 1}}{\frac{20.80 - 12.96}{105 - 2 - 1}} = \frac{0.322}{0.077} = 4.19 \tag{10.8.13}$$

The degrees of freedom are 1 and 102. The F distribution calculator shows that $p = 0.0432$.

The significance test of the variance explained uniquely by a variable is identical to a significance test of the regression coefficient for that variable. A regression coefficient and the variance explained uniquely by a variable both reflect the relationship between a variable and the criterion independent of the other variables. If the variance explained uniquely by a variable is not zero, then the regression coefficient cannot be zero. Clearly, a variable with a regression coefficient of zero would explain no variance.

Other inferential statistics associated with multiple regression are beyond the scope of this text. Two of particular importance are:

1. confidence intervals on regression slopes and
2. confidence intervals on predictions for specific observations.

These inferential statistics can be computed by standard statistical analysis packages such as R, $SPSS$, $STATA$, SAS, and JMP.

SPSS Output JMP Output

Assumptions

No assumptions are necessary for computing the regression coefficients or for partitioning the sum of squares. However, there are several assumptions made when interpreting inferential statistics. Moderate violations of Assumptions $1 - 3$ do not pose a serious

problem for testing the significance of predictor variables. However, even small violations of these assumptions pose problems for confidence intervals on predictions for specific observations.

1. **Residuals are normally distributed:**

   As in the case of simple linear regression, the residuals are the errors of prediction. Specifically, they are the differences between the actual scores on the criterion and the predicted scores. A $Q-Q$ plot for the residuals for the example data is shown below. This plot reveals that the actual data values at the lower end of the distribution do not increase as much as would be expected for a normal distribution. It also reveals that the highest value in the data is higher than would be expected for the highest value in a sample of this size from a normal distribution. Nonetheless, the distribution does not deviate greatly from normality.

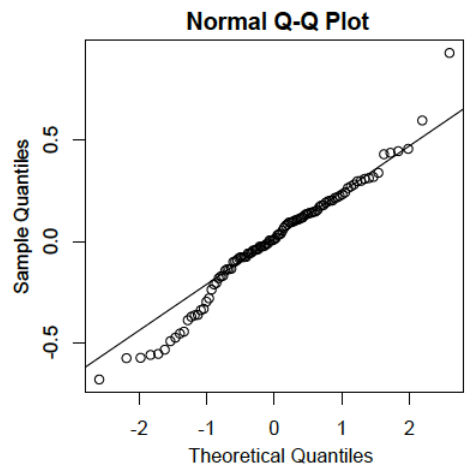


Figure 10.8.1: A $Q-Q$ plot for the residuals for the example data

2. **Homoscedasticity:**

   It is assumed that the variances of the errors of prediction are the same for all predicted values. As can be seen below, this assumption is violated in the example data because the errors of prediction are much larger for observations with low-to-medium predicted scores than for observations with high predicted scores. Clearly, a confidence interval on a low predicted $UGPA$ would underestimate the uncertainty.

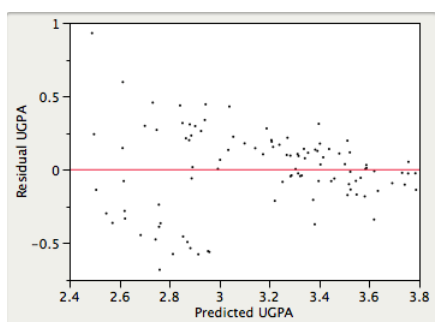


Figure 10.8.2: Low predicted $UGPA$

3. **Linearity:**

   It is assumed that the relationship between each predictor variable and the criterion variable is linear. If this assumption is not met, then the predictions may systematically overestimate the actual values for one range of values on a predictor variable and underestimate them for another.

10.9: Statistical Literacy

Learning Objectives

- Regression Toward the Mean in American Football

In a discussion about the Dallas Cowboy football team, there was a comment that the quarterback threw far more interceptions in the first two games than is typical (there were two interceptions per game). The author correctly pointed out that, because of regression toward the mean, performance in the future is expected to improve. However, the author defined regression toward the mean as, "In nerd land, that basically means that things tend to even out over the long run."

Example 10.9.1: what do you think?

Comment on that definition.

Solution

That definition is sort of correct, but it could be stated more precisely. Things don't always tend to even out in the long run. If a great player has an average game, then things wouldn't even out (to the average of all players) but would regress toward that player's high mean performance.

10.E: Regression (Exercises)

General Questions

Q1

What is the equation for a regression line? What does each term in the line refer to? (relevant section)

Q2

The formula for a regression equation based on a sample size of 25 observations is $Y' = 2X + 9$.

a. What would be the predicted score for a person scoring 6 on X?
b. If someone's predicted score was 14, what was this person's score on X? (relevant section)

Q3

What criterion is used for deciding which regression line fits best? (relevant section)

Q4

What does the standard error of the estimate measure? What is the formula for the standard error of the estimate? (relevant section)

Q5

a. In a regression analysis, the sum of squares for the predicted scores is 100 and the sum of squares error is 200, what is R^2?
b. In a different regression analysis, 40% of the variance was explained. The sum of squares total is 1000. What is the sum of squares of the predicted values? (relevant section)

Q6

For the X, Y data below, compute:

a. r and determine if it is significantly different from zero.
b. the slope of the regression line and test if it differs significantly from zero.
c. the 95% confidence interval for the slope.

(relevant section)

X	Y
2	5
4	6
4	7
5	11
6	12

Q7

What assumptions are needed to calculate the various inferential statistics of linear regression? (relevant section)

Q8

The correlation between years of education and salary in a sample of 20 people from a certain company is 0.4. Is this correlation statistically significant at the 0.05 level? (relevant section)

Q9

A sample of X and Y scores is taken, and a regression line is used to predict Y from X. If $SSY' = 300$, $SSE = 500$, and $N = 50$, what is: (relevant section relevant section)

a. SSY?
b. the standard error of the estimate?
c. R^2?

Q10

Using linear regression, find the predicted post-test score for someone with a score of 43 on the pre-test. (relevant section)

Pre	Post
59	56
52	63
44	55
51	50
42	66
42	48
41	58
45	36
27	13
63	50
54	81
44	56
50	64
47	50
55	63
49	57
45	73
57	63
46	46
60	60
65	47
64	73
50	58
74	85
59	44

Q11

The equation for a regression line predicting the number of hours of TV watched by children (Y) from the number of hours of TV watched by their parents (X) is $Y' = 4 + 1.2X$. The sample size is 12.

a. If the standard error of b is 0.4, is the slope statistically significant at the 0.05 level? (relevant section)

b. If the mean of X is 8, what is the mean of Y? (relevant section)

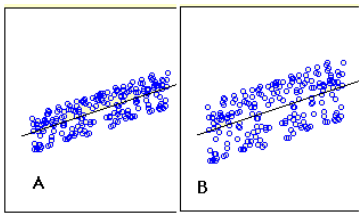
Q12

Based on the table below, compute the regression line that predicts Y from X. (relevant section)

MX	MY	sX	sY	r
10	12	2.5	3.0	-0.6

Q13

Does A or B have a larger standard error of the estimate? (relevant section)



Q14

True/false: If the slope of a simple linear regression line is statistically significant, then the correlation will also always be significant. (relevant section)

Q15

True/false: If the slope of the relationship between X and Y is larger for $\text{Population } 1$ than for $\text{Population } 2$, the correlation will necessarily be larger in $\text{Population } 1$ than in $\text{Population } 1$. Why or why not? (relevant section)

Q16

True/false: If the correlation is 0.8, then 40% of the variance is explained. (relevant section)

Q17

True/false: If the actual Y score was 31, but the predicted score was 28, then the error of prediction is 3. (relevant section)

Questions from Case Studies

The following question is from the Angry Moods (AM) case study.

Q18

(AM#23) Find the regression line for predicting Anger-Out from Control-Out.

a. What is the slope?
b. What is the intercept?
c. Is the relationship at least approximately linear?
d. Test to see if the slope is significantly different from 0.
e. What is the standard error of the estimate?

(relevant section, relevant section, relevant section)

The following question is from the SAT and GPA (SG) case study.

Q19

(SG#3) Find the regression line for predicting the overall university GPA from the high school GPA.

a. What is the slope?
b. What is the y-intercept?
c. If someone had a 2.2 GPA in high school, what is the best estimate of his or her college GPA?
d. If someone had a 4.0 GPA in high school, what is the best estimate of his or her college GPA?

(relevant section)

The following questions are from the Driving (D) case study.

Q20

(D#5) What is the correlation between age and how often the person chooses to drive in inclement weather? Is this correlation statistically significant at the 0.01 level? Are older people more or less likely to report that they drive in inclement weather? (relevant section, relevant section)

Q21

(D#8) What is the correlation between how often a person chooses to drive in inclement weather and the percentage of accidents the person believes occur in inclement weather? Is this correlation significantly different from 0? (relevant section, relevant section)

Q22

(D#10) Use linear regression to predict how often someone rides public transportation in inclement weather from what percentage of accidents that person thinks occur in inclement weather. (Pubtran by Accident)

a. Create a scatter plot of this data and add a regression line.
b. What is the slope?
c. What is the intercept?
d. Is the relationship at least approximately linear?
e. Test if the slope is significantly different from 0.
f. Comment on possible assumption violations for the test of the slope.
g. What is the standard error of the estimate?

(relevant section, relevant section, relevant section)

Selected Answers

S2

a. 21

S5

a. 0.33

S6

b. $b = 1.91$

S9

a. 800

S12

$a = 19.2$

S18

e. 3.45

S19

c. 2.6

S20

$r = 0.43$

S22

b. 0.35

CHAPTER OVERVIEW

11: Power

Power is defined as the probability of correctly rejecting a false null hypothesis. For example, it can be the probability that given there is a difference between the population means of the new method and the standard method, the sample means will be significantly different. The probability of failing to reject a false null hypothesis is often referred to as β. Therefore power can be defined as:

$$\text{power}=1-\beta$$

It is very important to consider power while designing an experiment. You should avoid spending a lot of time and/or money on an experiment that has little chance of finding a significant effect.

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

11.1: Introduction to Power

Learning Objectives

- Identify situations in which it is important to estimate power

Suppose you work for a foundation whose mission is to support researchers in mathematics education and your role is to evaluate grant proposals and decide which ones to fund. You receive a proposal to evaluate a new method of teaching high-school algebra. The research plan is to compare the achievement of students taught by the new method with the achievement of students taught by the traditional method. The proposal contains good theoretical arguments why the new method should be superior and the proposed methodology is sound. In addition to these positive elements, there is one important question still to be answered: Does the experiment have a high probability of providing strong evidence that the new method is better than the standard method if, in fact, the new method is actually better? It is possible, for example, that the proposed sample size is so small that even a fairly large population difference would be difficult to detect. That is, if the sample size is small, then even a fairly large difference in sample means might not be significant. If the difference is not significant, then no strong conclusions can be drawn about the population means. It is not justified to conclude that the null hypothesis that the population means are equal is true just because the difference is not significant. Of course, it is not justified to conclude that this null hypothesis is false. Therefore, when an effect is not significant, the result is inconclusive. You may prefer that your foundation's money be used to fund a project that has a higher probability of being able to make a strong conclusion.

Power is defined as the probability of correctly rejecting a false null hypothesis. In terms of our example, it is the probability that given there is a difference between the population means of the new method and the standard method, the sample means will be significantly different. The probability of failing to reject a false null hypothesis is often referred to as β. Therefore power can be defined as:

$$\text{power} = 1 - \beta \tag{11.1.1}$$

It is very important to consider power while designing an experiment. You should avoid spending a lot of time and/or money on an experiment that has little chance of finding a significant effect.

11.2: Example Calculations

skills to develop

- Compute power using the binomial distribution
- Compute power using the normal distribution
- Use a power calculator to compute power for the t distribution

In the "Shaking and Stirring Martinis" case study, the question was whether Mr. Bond could tell the difference between martinis that were stirred and martinis that were shaken. For the sake of this example, assume he can tell the difference and is able to correctly state whether a martini had been shaken or stirred 0.75 of the time. Now, suppose an experiment is being conducted to investigate whether Mr. Bond can tell the difference. Specifically, is Mr. Bond correct more than 0.50 of the time? We know that he is (that's an assumption of the example). However, the experimenter does not know and asks Mr. Bond to judge 16 martinis. The experimenter will do a significance test based on the binomial distribution. Specifically, if a one tailed test is significant at the 0.05 level, then he or she will conclude that Mr. Bond can tell the difference. The probability value is computed assuming the null hypothesis is true ($\pi = 0.50$). Therefore, the experimenter will determine how many times Mr. Bond is correct, and compute the probability of being correct that many or more times given that the null hypothesis is true. The question is: what is the probability the experimenter will correctly reject the null hypothesis that $\pi = 0.50$? In other words, what is the power of this experiment?

The binomial distribution for $N = 16$ and $\pi = 0.50$ is shown in Figure $11.2.1$. The probability of being correct on 11 or more trials is 0.105 and the probability of being correct on 12 or more trials is 0.038. Therefore, the probability of being correct on 12 or more trials is less than 0.05. This means that the null hypothesis will be rejected if Mr. Bond is correct on 12 or more trials and will not be rejected otherwise.

Binomial Calculator

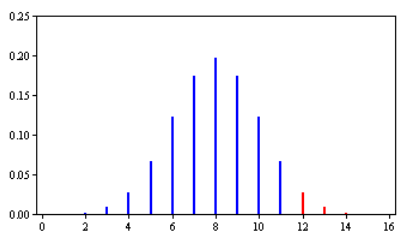


Figure 11.2.1: The binomial distribution for $N = 16$ and $\pi = 0.50$

We know that Mr. Bond is correct 0.75 of the time. (Obviously the experimenter does not know this or there would be no need for an experiment.) The binomial distribution with $N = 16$ and $\pi = 0.75$ is shown in Figure $11.2.2$

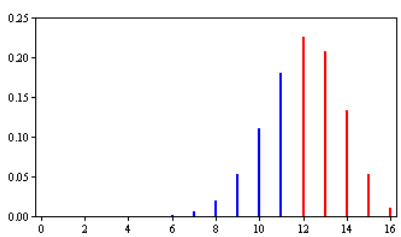


Figure 11.2.2: The binomial distribution for $N = 16$ and $\pi = 0.75$

The probability of being correct on 12 or more trials is 0.63. Therefore, the power of the experiment is 0.63.

To sum up, the probability of being correct on 12 or more trials given that the null hypothesis is true is less than 0.05. Therefore, if Mr. Bond is correct on 12 or more trials, the null hypothesis will be rejected. Given Mr. Bond's true ability to be correct on 0.75 of the trials, the probability he will be correct on 12 or more trials is 0.63. Therefore power is 0.63.

In the section on Testing a Single Mean for significance, the first example was based on the assumption that the experimenter knew the population variance. Although this is rarely true in practice, the example is very useful for pedagogical purposes. For the same reason, the following example assumes the experimenter knows the population variance. Power calculators are available for situations in which the experimenter does not know the population variance.

Suppose a math achievement test were known to have a mean of 75 and a standard deviation of 10. A researcher is interested in whether a new method of teaching results in a higher mean. Assume that although the experimenter does not know it, the population mean for the new method is 80. The researcher plans to sample 25 subjects and do a one-tailed test of whether the sample mean is significantly higher than 75. What is the probability that the researcher will correctly reject the false null hypothesis that the population mean for the new method is 75 or lower? The following shows how this probability is computed.

The researcher assumes that the population standard deviation with the new method is the same as with the old method (10) and that the distribution is normal. Since the population standard deviation is assumed to be known, the researcher can use the normal distribution rather than the t distribution to compute the p value. Recall that the standard error of the mean (σ_M) is

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{11.2.1}$$

which is equal to $10/5 = 2$ in this example. As can be seen in Figure $11.2.3$, if the null hypothesis that the population mean equals 75 is true, then the probability of a sample mean being greater than or equal to 78.29 is 0.05. Therefore, the experimenter will reject the null hypothesis if the sample mean, M, is 78.29 or larger.

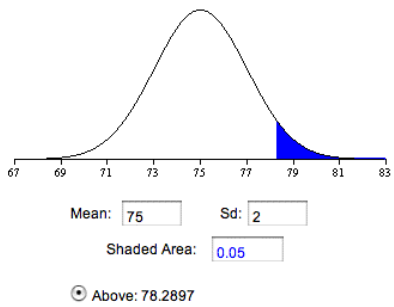


Figure created with the Inverse Normal Calculator)

The question, then, is what is the probability the experimenter gets a sample mean greater than 78.29 given that the population mean is 80? Figure $11.2.4$ shows that this probability is 0.80.

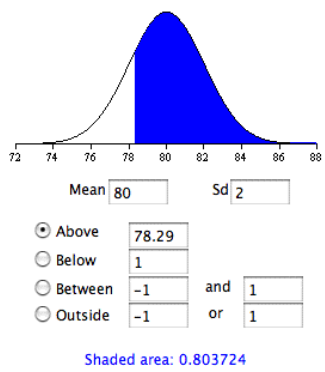


Figure created with the Normal Calculator)

Therefore, the probability that the experimenter will reject the null hypothesis that the population mean of the new method is 75 is 0.80. In other words, $power = 0.80$.

Calculation of power is more complex for t tests and for Analysis of Variance. The power calculator computes power for a t test of independent groups. Calculators for other types of designs are linked to below:

Russ Lenth's Power Calculators

11.3: Power Demo

Learning Objectives

- State the effect of one- versus two-tailed tests on power
- State the effect of α level on power

Instructions

This simulation illustrates the effect of

a. sample size
b. the difference between population mean and hypothesized mean
c. the standard deviation
d. the type of test (one-tailed or two)
e. significance level on the power of a two-sample t test

You specify the difference between the population mean and hypothesized mean by either entering text in the box or by moving the slider. The population standard deviation can be entered in the box or specified by the pop-up menu. Finally, you specify the number of tails of the test. A power graph for the significance level 0.10, 0.05, and 0.01 significance levels as a function of sample size is displayed.

1. Examine the power curves. The X axis shows sample size, the Y axis shows power. Note the effect of sample size and significance level on power.
2. The default difference between the population mean and hypothesized mean is 1.55. Use the slider to change this value and note the effect on the power curves.
3. Change the population standard deviation (sd) and notice its effect on power.
4. Compare the power of one-tailed and two-tailed tests when the difference between the population mean and hypothesized mean is 2. Determine which is higher.
5. Set the difference between the population mean and hypothesized mean to zero. Since the null hypothesis is true, the Y axis no shows the **Type I** error rate, not power. Note the effect of sample size on the **Type I** error rate.
6. Determine the effect of changing the standard deviation on the **Type I** error rate.

Illustrated Instructions

Video Demo

The video demonstration begins by changing the population mean and hypothesized mean difference to 2 and then the population standard deviation to 3.5. Notice how the power curves change with each adjustment. The video by switching between one-tailed and two tailed tests.

11.4: Power Demo II

Learning Objectives

- State the effect of effect size on power
- State the effect of one- versus two-tailed tests on power
- State the effect of the standard deviation on power
- State the effect of α level on power

Instructions

This demonstration shows various aspects of power. The graph displays power for a one-sample Z-test of the null hypothesis that the population mean is 50. The red distribution is the sampling distribution of the mean assuming the null hypothesis is true. The blue distribution is the sampling distribution of the mean based on the "true" mean. The default value of the true mean is 70. The cutoff for significance is based on the red distribution. For a one-tailed test at the 0.05 level, the cutoff point is determined so that 5% of the area is to the right of the cutoff. For the default values, the cutoff is 58.22. Therefore, any $\text{sample mean} = 58.22$ would be significant.

The shaded area of the blue distribution shows power. It is the probability that the sample mean will be 58.22 if the true mean is 70. For the default values it is 0.991.

You specify the true mean by either entering text in the box. The population standard deviation can be specified by the pop-up menu. Finally, you can specify the number of tails of the test and the significance level. Power and the cut-off points are displayed.

1. Notice the effect on power of changing the true mean to 60. Look at the distributions to find the area that represents power.
2. Change the sample size and notice its effect on the cutoff point and on power.
3. Change the population standard deviation (sd) and notice its effect on the cutoff point and on power.
4. Compare the power of one-tailed and two-tailed tests. What is the power for each when the true mean is 55. What about 45? (The one-tailed tests are based on the expectation that the $\text{population mean is} > 50$.)
5. Set the value of the true mean to 50. Does power make sense in this situation?
6. Determine the effect of the significance level on the red and blue distributions. On what (in addition to power) does it have its effect?

Illustrated Instructions

Video Demo

The video demonstration starts by changing the true mean to 55 and then the standard to 28. Notice how the probability of rejecting the null hypothesis varies with these changes.

The video concludes by changing the test to two-tailed and decreasing the significance level to 0.01. Again notice the changes in the probability of rejection.

11.5: Factors Affecting Power

Learning Objectives

- State what the effect of each of the factors is

Several factors affect the power of a statistical test. Some of the factors are under the control of the experimenter, whereas others are not. The following example will be used to illustrate the various factors.

Suppose a math achievement test were known to be normally distributed with a mean of 75 and a standard deviation of σ. A researcher is interested in whether a new method of teaching results in a higher mean. Assume that although the experimenter does not know it, the population mean μ for the new method is larger than 75. The researcher plans to sample N subjects and do a one-tailed test of whether the sample mean is significantly higher than 75. In this section, we consider factors that affect the probability that the researcher will correctly reject the false null hypothesis that the population mean is 75 or lower. In other words, factors that affect power.

Sample Size

Figure $11.5.1$ shows that the larger the sample size, the higher the power. Since sample size is typically under an experimenter's control, increasing sample size is one way to increase power. However, it is sometimes difficult and/or expensive to use a large sample size.

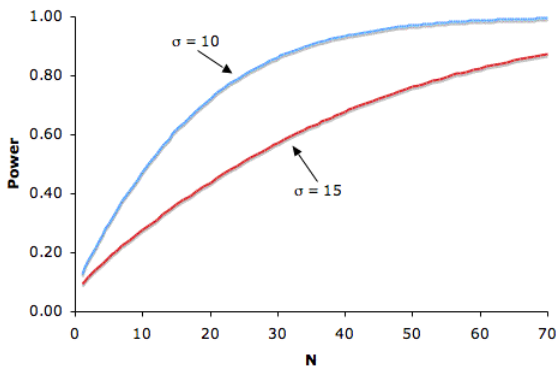


Figure 11.5.1: The relationship between sample size and power for $H_0 : \mu = 75$, real $\mu = 80$, one-tailed $\alpha = 0.05$, for $\sigma's$ of 10 and 15

Standard Deviation

Figure $11.5.1$ also shows that power is higher when the standard deviation is small than when it is large. For all values of N, power is higher for the standard deviation of 10 than for the standard deviation of 15 (except, of course, when $N = 0$). Experimenters can sometimes control the standard deviation by sampling from a homogeneous population of subjects, by reducing random measurement error, and/or by making sure the experimental procedures are applied very consistently.

Difference between Hypothesized and True Mean

Naturally, the larger the effect size, the more likely it is that an experiment would find a significant effect. Figure $11.5.2$ shows the effect of increasing the difference between the mean specified by the null hypothesis (75) and the population mean μ for standard deviations of 10 and 15.

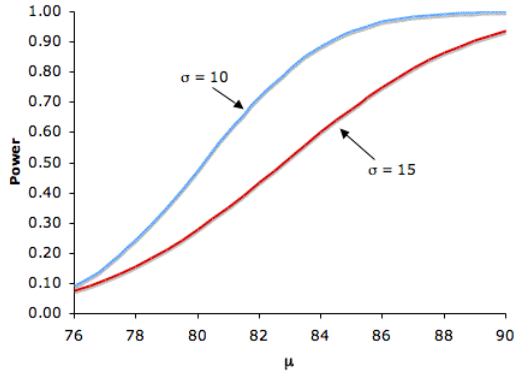


Figure 11.5.2: The relationship between μ and power for $H_0 : \mu = 75$, one-tailed $\alpha = 0.05$, for $\sigma' s$ of 10 and 15

Significance Level

There is a trade-off between the significance level and power: the more stringent (lower) the significance level, the lower the power. Figure $11.5.3$ shows that power is lower for the 0.01 level than it is for the 0.05 level. Naturally, the stronger the evidence needed to reject the null hypothesis, the lower the chance that the null hypothesis will be rejected.

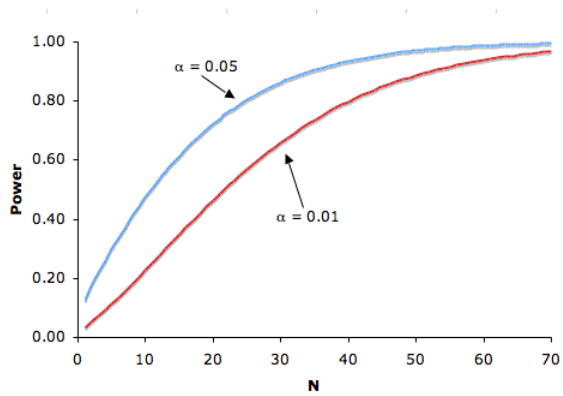


Figure $11.5.3$: The relationship between significance level and power with one-tailed tests: $\mu = 75$, real $\mu = 80$, and $\sigma = 10$

One- versus Two-Tailed Tests

Power is higher with a one-tailed test than with a two-tailed test as long as the hypothesized direction is correct. A one-tailed test at the 0.05 level has the same power as a two-tailed test at the 0.10 level. A one-tailed test, in effect, raises the significance level.

11.6: Statistical Literacy

Learning Objectives

- Design of Studies for Alzheimer's Drug

A research design to compare three drugs for the treatment of Alzheimer's disease is described here. For the first two years of the study, researchers will follow the subjects with scans and memory tests.

Example 11.6.1: what do you think?

The data could be analyzed as a between-subjects design or as a within-subjects design. What type of analysis would be done for each type of design and how would the choice of designs affect power?

Solution

For a between-subjects design, the subjects in the different conditions would be compared after two years. For a within-subjects design, the change in subjects' scores in the different conditions would be compared. The latter would be more powerful.

11.E: Power (Exercises)

General Questions

Q1

Define power in your own words.

Q2

List 3 measures one can take to increase the power of an experiment. Explain why your measures result in greater power.

Q3

$\text{Population 1 mean} = 36$

$\text{Population 2 mean} = 45$

Both population standard deviations are 10.

Sample size (per group) 16.

What is the probability that a t test will find a significant difference between means at the 0.05 level? Give results for both one- and two-tailed tests. Hint: the power of a one-tailed test at 0.05 level is the power of a two-tailed test at 0.10.

Q4

Rank order the following in terms of power. n is the sample size per group.

	Population 1 Mean	n	Population 2 Mean	Standard Deviation
a	29	20	43	12
b	34	15	40	6
c	105	24	50	27
d	170	2	120	10

Q5

Alan, while snooping around his grandmother's basement stumbled upon a shiny object protruding from under a stack of boxes . When he reached for the object a genie miraculously materialized and stated: "You have found my magic coin. If you flip this coin an infinite number of times you will notice that heads will show 60% of the time." Soon after the genie's declaration he vanished, never to be seen again. Alan, excited about his new magical discovery, approached his friend Ken and told him about what he had found. Ken was skeptical of his friend's story, however, he told Alan to flip the coin 100 times and to record how many flips resulted with heads.

a. What is the probability that Alan will be able convince Ken that his coin has special powers by finding a p value below 0.05 (one tailed). Use the Binomial Calculator (and some trial and error).
b. If Ken told Alan to flip the coin only 20 times, what is the probability that Alan will not be able to convince Ken (by failing to reject the null hypothesis at the 0.05 level)?

CHAPTER OVERVIEW

12: Effect Size

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

12.1: Prelude to Effect Size

Researchers often seek to learn more than whether the variable under investigation has an effect and/or the direction of the effect. This is particularly true for research that has practical applications. For example, an investigation of the efficacy of a pain-relief drug would seek to determine the extent of the relief and not merely whether there was any relief. Similarly, a study of a test-preparation course's efficacy would seek to determine how much the course raises students' test scores. Finally, a study of the relationship between exercise and blood pressure would seek to determine how much blood pressure decreases for a given amount of exercise. In all of these examples, a significance test would not be sufficient since it would only provide the researcher with information about the existence and direction of the effect. It would not provide any information about the size of the effect.

Before we proceed with a discussion of how to measure effect size, it is important to consider that for some research it is the presence or absence of an effect rather than its size that is important. A controversial example is provided by Bem (2011) who investigated precognition. Bem found statistically significant evidence that subjects' responses are affected by future events. That is, he rejected the null hypothesis that there is no effect. The important question is not the size of the effect but, rather, whether it exists at all. It would be truly remarkable if future events affect present responses even a little. It is important to note that subsequent research (Ritchie, Wiseman, & French, 2012) has failed to replicate Bem's results and the likelihood that the precognition effects he described are real is very low.

Bem, D. J. (201). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Journal of Personality and Social Psychology, 100, 407–425.

Ritchie, S. J., Wiseman R., and French, C. C. (2012) Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. PLoS ONE 7.

12.2: Proportions

Learning Objectives

- Compute absolute risk reduction
- Compute relative risk reduction
- Compute number needed to treat

Often the interpretation of a proportion is self-evident. For example, the obesity rate for white non-Hispanic adults living in the United States was estimated by a study conducted between 2006 and 2008 to be 24%. This value of 24% is easily interpretable and indicates the magnitude of the obesity problem in this population.

Often the question of interest involves the comparison of two outcomes. For example, consider the analysis of proportions in the case study "Mediterranean Diet and Health." In this study, one group of people followed the diet recommended by the American Heart Association (AHA), whereas a second group followed the "Mediterranean Diet." One interesting comparison is between the proportions of people who were healthy throughout the study as a function of diet. It turned out that 0.79 of the people who followed the AHA diet and 0.90 of those who followed the Mediterranean diet were healthy. How is the effect size of diet best measured?

We will take the perspective that we are assessing the benefits of switching from the AHA diet to the Mediterranean diet. One way to assess the benefits is to compute the difference between the proportion who were not healthy on the AHA diet (0.21) with the proportion who were not healthy on the Mediterranean diet (0.10). Therefore, the difference in proportions is:

$$0.21 - 0.10 = 0.11 \tag{12.2.1}$$

This measure of the benefit is called the **Absolute Risk Reduction** (ARR).

To define ARR more formally, let C be the proportion of people in the control group with the ailment of interest and T be the proportion in the treatment group. ARR can then be defined as:

$$ARR = C - T \tag{12.2.2}$$

Alternatively, one could measure the difference in terms of percentages. For our example, the proportion of non-healthy people on the Mediterranean diet (0.10) is 52% lower than the proportion of non-healthy people on the AHA diet (0.21). This value is computed as follows:

$$\frac{(0.21 - 0.10)}{0.21} \times 100 = 52\% \tag{12.2.3}$$

This measure of the benefit is called the **Relative Risk Reduction** (RRR). The general formula for RRR is:

$$RRR = \frac{(C - T)}{C} \times 100 \tag{12.2.4}$$

where C and T are defined as before.

A third commonly used measure is the "*odds ratio*." For our example, the odds of being healthy on the Mediterranean diet are $90 : 10 = 9 : 1$; the odds on the AHA diet are $79 : 21 = 3.76 : 1$. The ratio of these two odds is $9/3.76 = 2.39$. Therefore, the odds of being healthy on the Mediterranean diet is 2.39 times the odds of being healthy on the AHA diet. Note that the odds ratio is the ratio of the odds and not the ratio of the probabilities.

A fourth measure is the number of people who need to be treated in order to prevent one person from having the ailment of interest. In our example, being treated means changing from the AHA diet to the Mediterranean diet. The number who need to be treated can be defined as:

$$N = \frac{1}{ARR} \tag{12.2.5}$$

For our example,

$$N = \frac{1}{0.11} = 9 \tag{12.2.6}$$

Therefore, one person who would otherwise not be healthy would be expected to stay healthy for every nine people changing from the AHA diet to the Mediterranean diet.

The obvious question is which of these measures is the best one. Although each measure has its proper uses, the RRR measure can exaggerate the importance of an effect, especially when the absolute risks are low. For example, if a drug reduced the risk of a certain disease from 1 in $1,000,000$ to 1 in $2,000,000$, the RRR is 50%. However, since the ARR is only 0.0000005, the practical reduction in risk is minimal.

12.3: Difference Between Two Means

Learning Objectives

- State how the inherent meaningfulness of the scales affects the type of measure that should be used
- Compute g
- Compute d
- State the effect of the variability of subjects on the size of standardized measures

When the units of a measurement scale are meaningful in their own right, then the difference between means is a good and easily interpretable measure of effect size. For example, a study conducted by Holbrook, Crowther, Lotter, Cheng and King in 2000 investigated the effectiveness of benzodiazepine for the treatment of insomnia. These researchers found that, compared to a placebo, this drug increased total sleep duration by a mean of 61.8 minutes. This difference in means shows clearly the degree to which benzodiazepine is effective. (It is important to note that the drug was found to sometimes have adverse side effects.)

When the dependent variable is measured on a ratio scale, it is often informative to consider the proportional difference between means in addition to the absolute difference. For example, if in the Holbrook et al. study the mean total sleep time for the placebo group was 120 minutes, then the 61.8-minute increase would represent a 51% increase in sleep time. On the other hand, if the mean sleep time for the placebo were 420 minutes, then the 61.8-minute increase would represent a 15% increase in sleep time.

It is interesting to note that if a log transformation is applied to the dependent variable, then equal percent changes on the original scale will result in equal absolute changes on the log scale. For example, suppose the mean sleep time increased 10% from 400 minutes to 440 in one condition and 10% from 300 to 330 minutes in a second condition. If we take the log base 10 of these values, we find that

$$\log_{10}(440) - \log_{10}(400) = 2.643 - 2.602 = 0.041 \tag{12.3.1}$$

Similarly,

$$\log_{10}(330) - \log_{10}(300) = 2.518 - 2.477 = 0.041 \tag{12.3.2}$$

Many times the dependent variable is measured on a scale that is not inherently meaningful. For example, in the "Animal Research" case study, attitudes toward animal research were measured on a 7-point scale. The mean rating of women on whether animal research is wrong was 1.47 scale units higher than the mean rating of men. However, it is not clear whether this 1.47-unit difference should be considered a large effect or a small effect, since it is not clear exactly what this difference means.

When the scale of a dependent variable is not inherently meaningful, it is common to consider the difference between means in standardized units. That is, effect size is measured in terms of the number of standard deviations the means differ by. Two commonly used measures are Hedges' g and Cohen's d. Both of these measures consist of the difference between means divided by the standard deviation. They differ only in that Hedges' g uses the version of the standard deviation formula in which you divide by $N-1$, whereas Cohen's d uses the version in which you divide by N. The two formulas are given below.

$$g = \frac{M_1 - M_2}{\sqrt{MSE}} \tag{12.3.3}$$

$$d = g\sqrt{\frac{N}{N-2}} \tag{12.3.4}$$

where M_1 is the mean of the first group, M_2 is the mean of the second group, MSE is the mean square error, and N is the total number of observations.

Standardized measures such as Cohen's d and Hedges' g have the advantage that they are scale free. That is, since the dependent variable is standardized, the original units are replaced by standardized units and are interpretable even if the original scale units do not have clear meaning. Consider the Animal Research case study in which attitudes were measured on a 7-point scale. On a rating of whether animal research is wrong, the mean for women was 5.353, the mean for men was 3.882, and MSE was 2.864. Hedges' g can be calculated to be 0.87. It is more meaningful to say that the means were 0.87 standard deviations apart than 1.47 scale units apart, since the scale units are not well-defined.

It is natural to ask what constitutes a large effect. Although there is no objective answer to this question, the guidelines suggested by Cohen (1988) stating that an effect size of 0.2 is a small effect, an effect size of 0.5 is a medium effect, and an effect size of 0.8 is a large effect have been widely adopted. Based on these guidelines, the effect size of 0.87 is a large effect.

It should be noted, however, that these guidelines are somewhat arbitrary and have not been universally accepted. For example, Lenth (2001) argued that other important factors are ignored if Cohen's definition of effect size is used to choose a sample size to achieve a given level of power.

Interpretational Issues

It is important to realize that the importance of an effect depends on the context. For example, a small effect can make a big difference if only extreme observations are of interest. Consider a situation in which a test is used to select students for a highly selective program. Assume that there are two types of students (red and blue) and that the mean for the red students is 52, the mean for the blue students is 50, both distributions are normal, and the standard deviation for each distribution is 10. The difference in means is therefore only 0.2 standard deviations and would generally be considered to be a small difference. Now assume that only students who scored 70 or higher would be selected for the program. Would there be a big difference between the proportion of blue and red students who would be able to be accepted into the program? It turns out that the proportion of red students who would qualify is 0.036 and the proportion of blue students is 0.023. Although this difference is small in absolute terms, the ratio of red to blue students who qualify is $1.6 : 1$. This means that if 100 students were to be accepted and if equal numbers of randomly-selected red and blue students applied, 62% would be red and 38% would be blue. In most contexts this would be considered an important difference.

When the effect size is measured in standard deviation units as it is for Hedges' g and Cohen's d, it is important to recognize that the variability in the subjects has a large influence on the effect size measure. Therefore, if two experiments both compared the same treatment to a control but the subjects were much more homogeneous in $\text{Experiment } 1$ than in $\text{Experiment } 2$, then a standardized effect size measure would be much larger in the former experiment than in the latter. Consider two hypothetical experiments on the effect of an exercise program on blood pressure. Assume that the mean effect on systolic blood pressure of the program is $10mm\ Hg$ and that, due to differences in the subject populations sampled in the two experiments, the standard deviation was 20 in $\text{Experiment } 1$ and 30 in $\text{Experiment } 2$. Under these conditions, the standardized measure of effect size would be 0.50 in $\text{Experiment } 1$ and 0.33 in $\text{Experiment } 2$. This standardized difference in effect size occurs even though the effectiveness of the treatment is exactly the same in the two experiments.

Reference

1. Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates.
2. Lenth, R. V. (2001) Some Practical Guidelines for Effective Sample Size Determination. The American Statistician, 55, 187-193.

12.4: Proportion of Variance Explained

Learning Objectives

- State the difference in bias between η^2 and ω^2
- Compute η^2 Compute ω^2
- Distinguish between ω^2 and partial ω^2
- State the bias in R^2 and what can be done to reduce it

Effect sizes are often measured in terms of the proportion of variance explained by a variable. In this section, we discuss this way to measure effect size in both ANOVA designs and in correlational studies.

ANOVA Designs

Responses of subjects will vary in just about every experiment. Consider, for example, the "Smiles and Leniency" case study. A histogram of the dependent variable "leniency" is shown in Figure $12.4.1$. It is clear that the leniency scores vary considerably. There are many reasons why the scores differ. One, of course, is that subjects were assigned to four different smile conditions and the condition they were in may have affected their leniency score. In addition, it is likely that some subjects are generally more lenient than others, thus contributing to the differences among scores. There are many other possible sources of differences in leniency ratings including, perhaps, that some subjects were in better moods than other subjects and/or that some subjects reacted more negatively than others to the looks or mannerisms of the stimulus person. You can imagine that there are innumerable other reasons why the scores of the subjects could differ.

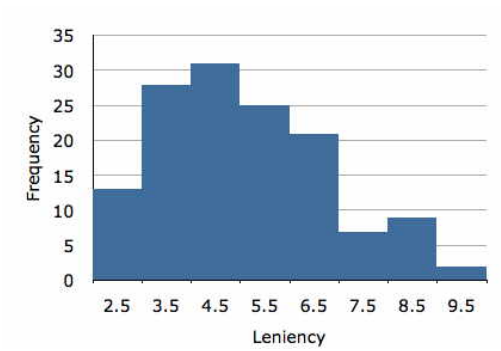


Figure 12.4.1: Distribution of leniency scores

One way to measure the effect of conditions is to determine the proportion of the variance among subjects' scores that is attributable to conditions. In this example, the variance of scores is 2.794. The question is how this variance compares with what the variance would have been if every subject had been in the same treatment condition. We estimate this by computing the variance within each of the treatment conditions and taking the mean of these variances. For this example, the mean of the variances is 2.649. Since the mean variance within the smile conditions is not that much less than the variance ignoring conditions, it is clear that "Smile Condition" is not responsible for a high percentage of the variance of the scores. The most convenient way to compute the proportion explained is in terms of the sum of squares "conditions" and the sum of squares total. The computations for these sums of squares are shown in the chapter on ANOVA. For the present data, the sum of squares for "Smile Condition" is 27.535 and the sum of squares total is 377.189. Therefore, the proportion explained by "Smile Condition" is:

$$\frac{27.535}{377.189} = 0.073 \tag{12.4.1}$$

Thus, 0.073 or 7.3% of the variance is explained by "Smile Condition."

An alternative way to look at the variance explained is as the proportion reduction in error. The sum of squares total (377.189) represents the variation when "Smile Condition" is ignored and the sum of squares error ($377.189 - 27.535 = 349.654$) is the variation left over when "Smile Condition" is accounted for. The difference between 377.189 and 349.654 is 27.535. This reduction in error of 27.535 represents a proportional reduction of $27.535/377.189 = 0.073$, the same value as computed in terms of proportion of variance explained.

This measure of effect size, whether computed in terms of variance explained or in terms of percent reduction in error, is called η^2 where η is the Greek letter eta. Unfortunately, η^2 tends to overestimate the variance explained and is therefore a biased estimate of the proportion of variance explained. As such, it is not recommended (despite the fact that it is reported by a leading statistics package).

An alternative measure, ω^2 (omega squared), is unbiased and can be computed from

$$\omega^2 = \frac{SSQ_{condition} - (k-1)MSE}{SSQ_{total} + MSE} \tag{12.4.2}$$

where MSE is the mean square error and k is the number of conditions. For this example, $k = 4$ and $\omega^2 = 0.052$.

It is important to be aware that both the variability of the population sampled and the specific levels of the independent variable are important determinants of the proportion of variance explained. Consider two possible designs of an experiment investigating the effect of alcohol consumption on driving ability. As can be seen in Table $12.4.1$, Design 1 has a smaller range of doses and a more diverse population than Design 2. What are the implications for the proportion of variance explained by Dose? Variation due to Dose would be greater in Design 2 than Design 1 since alcohol is manipulated more strongly than in Design 1. However, the variance in the population should be greater in Design 1 since it includes a more diverse set of drivers. Since with Design 1 the variance due to Dose would be smaller and the total variance would be larger, the proportion of variance explained by Dose would be much less using Design 1 than using Design 2. Thus, the proportion of variance explained is not a general characteristic of the independent variable. Instead, it is dependent on the specific levels of the independent variable used in the experiment and the variability of the population sampled.

Table 12.4.1: Design Parameters

Design	Dose	Population
1	0.00	All Drivers between 16 and 80 Years of Age
	0.30	
	0.60	
2	0.00	Experienced Drivers between 25 and 30 Years of Age
	0.50	
	1.00	

Factorial Designs

In one-factor designs, the sum of squares total is the sum of squares condition plus the sum of squares error. The proportion of variance explained is defined relative to sum of squares total. In an $A \times B$ design, there are three sources of variation ($A, B, A \times B$) in addition to error. The proportion of variance explained for a variable (A, for example) could be defined relative to the sum of squares total ($SSQ_A + SSQ_B + SSQ_{A \times B} + SSQ_{error}$) or relative to $SSQ_A + SSQ_{error}$.

To illustrate with an example, consider a hypothetical experiment on the effects of age (6 and 12 years) and of methods for teaching reading (experimental and control conditions). The means are shown in Table $12.4.2$ The standard deviation of each of the four cells ($Age \times Treatment$ combinations) is 5. (Naturally, for real data, the standard deviations would not be exactly equal and the means would not be whole numbers.) Finally, there were 10 subjects per cell resulting in a total of 40 subjects.

Table 12.4.2: Condition Means

	Treatment	
Age	Experimental	Control
6	40	42
12	50	56

The sources of variation, degrees of freedom, and sums of squares from the analysis of variance summary table as well as four measures of effect size are shown in Table 12.4.3. Note that the sum of squares for age is very large relative to the other two effects. This is what would be expected since the difference in reading ability between 6- and 12-year-olds is very large relative to the effect of condition.

Table 12.4.3: ANOVA Summary Table

Source	df	SSQ	η^2	partial η^2	ω^2	partial ω^2
Age	1	1440	0.567	0.615	0.552	0.586
Condition	1	160	0.063	0.151	0.053	0.119
A x C	1	40	0.016	0.043	0.006	0.015
Error	36	900				
Total	39	2540				

First, we consider the two methods of computing η^2, labeled η^2 and partial η^2. The value of η^2 for an effect is simply the sum of squares for this effect divided by the sum of squares total. For example, the η^2 for Age is $1440/2540 = 0.567$ As in a one-factor design, η^2 is the proportion of the total variation explained by a variable. Partial η^2 for Age is SSQ_{Age} divided by ($SSQ_{Age} + SSQ_{error}$), which is $1440/2340 = 0.615$

As you can see, the partial η^2 is larger than η^2. This is because the denominator is smaller for the partial η^2. The difference between η^2 and partial η^2 is even larger for the effect of condition. This is because SSQ_{Age} is large and it makes a big difference whether or not it is included in the denominator.

As noted previously, it is better to use ω^2 than η^2 because η^2 has a positive bias. You can see that the values for ω^2 are smaller than for η^2. The calculations for ω^2 are shown below:

$$\omega^2 = \frac{SSQ_{effect} - df_{effect} MS_{error}}{SSQ_{total} + MS_{error}} \tag{12.4.3}$$

$$\omega^2_{partial} = \frac{SSQ_{effect} - df_{effect} MS_{error}}{SSQ_{effect} + (N - df_{effect}) MS_{error}} \tag{12.4.4}$$

where N is the total number of observations.

The choice of whether to use ω^2 or the partial ω^2 is subjective; neither one is correct or incorrect. However, it is important to understand the difference and, if you are using computer software, to know which version is being computed. (Beware, at least one software package labels the statistics incorrectly).

Correlational Studies

In the section "Partitioning the Sums of Squares" in the Regression chapter, we saw that the sum of squares for Y (the criterion variable) can be partitioned into the sum of squares explained and the sum of squares error. The proportion of variance explained in

multiple regression is therefore:

$$SSQ_{explained} / SSQ_{total} \tag{12.4.5}$$

In simple regression, the proportion of variance explained is equal to r^2; in multiple regression, it is equal to R^2.

In general, R^2 is analogous to η^2 and is a biased estimate of the variance explained. The following formula for adjusted R^2 is analogous to ω^2 and is less biased (although not completely unbiased):

$$R^2_{adjusted} = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \tag{12.4.6}$$

where N is the total number of observations and p is the number of predictor variables.

12.5: Statistical Literacy

Learning Objectives

- Health Effects of Coffee

This article describes some health effects of drinking coffee. Among the key findings were

a. women who drank four or more cups a day reduced their risk of endometrial cancer by 25% compared with those who drank less than one cup a day
b. men who drank six or more cups had a 60% lower risk of developing the most deadly form of prostate cancer than those who drank less than one cup a day

Example 12.5.1: what do you think?

What is the technical term for the measure of risk reduction reported? What measures of risk reduction cannot be determined from the article? What additional information would have been helpful for assessing risk reduction?

Solution

This is called the "relative risk reduction." The article does not provide information necessary to compute the absolute risk reduction, the odds ratio, or the number needed to treat. It would have been helpful if the article had reported the proportion of women drinking less than one cup a day who developed endometrial cancer as well as the analogous statistic for men and prostate cancer.

12.6: Effect Size (Exercises)

General Questions

Q1

If the probability of a disease is 0.34 without treatment and 0.22 with treatment then what is the

a. Absolute risk reduction
b. Relative risk reduction
c. Odds ratio
d. Number needed to treat

Q2

When is it meaningful to compute the proportional difference between means?

Q3

The mean for an experimental group is 12, the mean for the control group were 8, the MSE from the ANOVA is 16, and N, the number of observations is 20, compute g and d.

Q4

Two experiments investigated the same variables but one of the experiment had subject who differed greatly from each other where as the subjects in the other experiment were relatively homogeneous. Which experiment would likely have the larger value of g?

Q5

Why is ω^2 preferable to η^2?

Q6

What is the difference between η^2 and partial η^2?

Questions from Case Studies

The following questions are from the Teacher Ratings case study.

Q7

What are the values of d and g?

Q8

What are the values of ω^2 and η^2?

The following question is from the Smiles and Leniency case study.

Q9

What are the values of ω^2 and η^2?

The following question is from the Obesity and Bias case study.

Q10

For compute ω^2 and partial ω^2 for the effect of "Weight" in a "Weight x Relatedness" ANOVA.

CHAPTER OVERVIEW

13: Analysis of Variance

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

13.1: Introduction to ANOVA

Learning Objectives

- What null hypothesis is tested by ANOVA
- Describe the uses of ANOVA

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called "Analysis of Variance" rather than "Analysis of Means." As you will see, the name is appropriate because inferences about means are made by analyzing variance.

ANOVA is used to test general rather than specific differences among means. This can be seen best by example. In the case study "Smiles and Leniency," the effect of different types of smiles on the leniency shown to a person was investigated. Four different types of smiles (neutral, false, felt, miserable) were investigated. The chapter "All Pairwise Comparisons among Means" showed how to test differences among means. The results from the Tukey HSD test are shown in Table 13.1.1.

Table 13.1.1: *Six Pairwise Comparisons*

Comparison	M_i-M_j	Q	p
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

Notice that the only significant difference is between the False and Neutral conditions.

ANOVA tests the non-specific null hypothesis that all four population means are equal. That is,

$$\mu_{false} = \mu_{felt} = \mu_{miserable} = \mu_{neutral} \tag{13.1.1}$$

This non-specific null hypothesis is sometimes called the omnibus null hypothesis. When the omnibus null hypothesis is rejected, the conclusion is that at least one population mean is different from at least one other mean. However, since the ANOVA does not reveal which means are different from which, it offers less specific information than the *Tukey HSD test*. The Tukey HSD is therefore preferable to ANOVA in this situation. Some textbooks introduce the Tukey test only as a follow-up to an ANOVA. However, there is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA.

You might be wondering why you should learn about ANOVA when the Tukey test is better. One reason is that there are complex types of analysis that can be done with ANOVA and not with the Tukey test. A second is that ANOVA is by far the most commonly-used technique for comparing means, and it is important to understand ANOVA in order to understand research reports.

13.2: ANOVA Designs

Learning Objectives

- Determine whether a factor is a between-subjects or a within-subjects factor
- Define factorial design

There are many types of experimental designs that can be analyzed by ANOVA. This section discusses many of these designs and defines several key terms used.

Factors and Levels

The section on variables defined an independent variable as a variable manipulated by the experimenter. In the case study "Smiles and Leniency," the effect of different types of smiles on the leniency shown to a person was investigated. Four different types of smiles (neutral, false, felt, and miserable) were shown. In this experiment, "Type of Smile" is the independent variable. In describing an ANOVA design, the term factor is a synonym of independent variable. Therefore, "Type of Smile" is the factor in this experiment. Since four types of smiles were compared, the factor "Type of Smile" has four levels.

An ANOVA conducted on a design in which there is only one factor is called a one-way ANOVA. If an experiment has two factors, then the ANOVA is called a two-way ANOVA. For example, suppose an experiment on the effects of age and gender on reading speed were conducted using three age groups (8 years, 10 years, and 12 years) and the two genders (male and female). The factors would be age and gender. Age would have three levels and gender would have two levels.

Between- and Within-Subjects Factors

In the "Smiles and Leniency" study, the four levels of the factor "Type of Smile" were represented by four separate groups of subjects. When different subjects are used for the levels of a factor, the factor is called a between-subjects factor or a between-subjects variable. The term "between subjects" reflects the fact that comparisons are between different groups of subjects.

In the "ADHD Treatment" study, every subject was tested with each of four dosage levels $(0, 0.15, 0.30, 0.60$mg/kg) of a drug. Therefore there was only one group of subjects, and comparisons were not between different groups of subjects but between conditions within the same subjects. When the same subjects are used for the levels of a factor, the factor is called a within-subjects factor or a within-subjects variable. Within-subjects variables are sometimes referred to as repeated-measures variables since there are repeated measurements of the same subjects.

Multi-Factor Designs

It is common for designs to have more than one factor. For example, consider a hypothetical study of the effects of age and gender on reading speed in which males and females from the age levels of 8 years, 10 years, and 12 years are tested. There would be a total of six different groups as shown in Table $13.2.1$.

Table $13.2.1$: Gender x Age Design

Group	Gender	Age
1	Female	8
2	Female	10
3	Female	12
4	Male	8
5	Male	10
6	Male	12

This design has two factors: age and gender. Age has three levels and gender has two levels. When all combinations of the levels are included (as they are here), the design is called a factorial design. A concise way of describing this design is as a $Gender(2) \times Age(3)$ factorial design where the numbers in parentheses indicate the number of levels. Complex designs frequently have more than two factors and may have combinations of between- and within-subjects factors.

13.3: One-Factor ANOVA

Learning Objectives

- State what the Mean Square Between (MSB) estimates when the null hypothesis is true and when the null hypothesis is false
- Compute MSE
- Compute F and its two degrees of freedom parameters
- Explain why ANOVA is best thought of as a two-tailed test even though literally only one tail of the distribution is used
- Partition the sums of squares into condition and error
- Format data to be used with a computer statistics program

This section shows how ANOVA can be used to analyze a one-factor between-subjects design. We will use as our main example the "Smiles and Leniency" case study. In this study there were four conditions with 34 subjects in each condition. There was one score per subject. The null hypothesis tested by ANOVA is that the population means for all conditions are the same. This can be expressed as follows:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \tag{13.3.1}$$

where H_0 is the null hypothesis and k is the number of conditions. In the "Smiles and Leniency" study, $k = 4$ and the null hypothesis is

$$H_0 : \mu_{false} = \mu_{felt} = \mu_{miserable} = \mu_{neutral} \tag{13.3.2}$$

If the null hypothesis is rejected, then it can be concluded that at least one of the population means is different from at least one other population mean.

Analysis of variance is a method for testing differences among means by analyzing variance. The test is based on two estimates of the population variance (σ^2). One estimate is called the mean square error (MSE) and is based on differences among scores within the groups. MSE estimates σ^2 regardless of whether the null hypothesis is true (the population means are equal). The second estimate is called the mean square between (MSB) and is based on differences among the sample means. MSB only estimates σ^2 if the population means are equal. If the population means are not equal, then MSB estimates a quantity larger than σ^2. Therefore, if the MSB is much larger than the MSE, then the population means are unlikely to be equal. On the other hand, if the MSB is about the same as MSE, then the data are consistent with the null hypothesis that the population means are equal.

Before proceeding with the calculation of MSE and MSB, it is important to consider the assumptions made by ANOVA:

1. The populations have the same variance. This assumption is called the assumption of $homogeneity\ of\ variance$.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the values are not independent. The analysis of data with two scores per subject is shown in the section on within-subjects ANOVA later in this chapter.

These assumptions are the same as for a t test of differences between groups except that they apply to two or more groups, not just to two groups.

The means and variances of the four groups in the "Smiles and Leniency" case study are shown in Table $13.3.1$. Note that there are 34 subjects in each of the four conditions (False, Felt, Miserable, and Neutral).

Table $13.3.1$: Means and Variances from the "Smiles and Leniency" Study

Condition	Mean	Variance
False	5.3676	3.3380
Felt	4.9118	2.8253
Miserable	4.9118	2.1132
Neutral	4.1176	2.3191

Sample Sizes

The first calculations in this section all assume that there is an equal number of observations in each group. Unequal sample size calculations are shown here. We will refer to the number of observations in each group as n and the total number of observations as N. For these data there are four groups of 34 observations. Therefore, $n = 34$ and $N = 136$.

Computing MSE

Recall that the assumption of homogeneity of variance states that the variance within each of the populations (σ^2) is the same. This variance, σ^2, is the quantity estimated by MSE and is computed as the mean of the sample variances. For these data, the MSE is equal to 2.6489.

Computing MSB

The formula for MSB is based on the fact that the variance of the sampling distribution of the mean is

$$\sigma_M^2 = \frac{\sigma^2}{n} \tag{13.3.3}$$

where n is the sample size of each group. Rearranging this formula, we have

$$\sigma^2 = n\sigma_M^2 \tag{13.3.4}$$

Therefore, if we knew the variance of the sampling distribution of the mean, we could compute σ^2 by multiplying it by n. Although we do not know the variance of the sampling distribution of the mean, we can estimate it with the variance of the sample means. For the leniency data, the variance of the four sample means is 0.270. To estimate σ^2, we multiply the variance of the sample means (0.270) by n (the number of observations in each group, which is 34). We find that $MSB = 9.179$.

To sum up these steps:

1. Compute the means.
2. Compute the variance of the means.
3. Multiply the variance of the means by n.

Recap

If the population means are equal, then both MSE and MSB are estimates of σ^2 and should therefore be about the same. Naturally, they will not be exactly the same since they are just estimates and are based on different aspects of the data: The MSB is computed from the sample means and the MSE is computed from the sample variances.

If the population means are not equal, then MSE will still estimate σ^2 because differences in population means do not affect variances. However, differences in population means affect MSB since differences among population means are associated with differences among sample means. It follows that the larger the differences among sample means, the larger the MSB.

Note

In short, MSE estimates σ^2 whether or not the population means are equal, whereas MSB estimates σ^2 only when the population means are equal and estimates a larger quantity when they are not equal.

Comparing MSE and MSB

The critical step in an ANOVA is comparing MSE and MSB. Since MSB estimates a larger quantity than MSE only when the population means are not equal, a finding of a larger MSB than an MSE is a sign that the population means are not equal. But since MSB could be larger than MSE by chance even if the population means are equal, MSB must be much larger than MSE in order to justify the conclusion that the population means differ. But how much larger must MSB be? For the "Smiles and Leniency" data, the MSB and MSE are 9.179 and 2.649, respectively. Is that difference big enough? To answer, we would need to know the probability of getting that big a difference or a bigger difference if the population means were all equal. The mathematics necessary to answer this question were worked out by the statistician R. Fisher. Although Fisher's original formulation took a slightly different form, the standard method for determining the probability is based on the ratio of MSB to MSE. This ratio is named after Fisher and is called the F ratio.

For these data, the F ratio is

$$F = \frac{9.179}{2.649} = 3.465 \tag{13.3.5}$$

Therefore, the MSB is 3.465 times higher than MSE. Would this have been likely to happen if all the population means were equal? That depends on the sample size. With a small sample size, it would not be too surprising because results from small samples are unstable. However, with a very large sample, the MSB and MSE are almost always about the same, and an F ratio of 3.465 or larger would be very unusual. Figure $13.3.1$ shows the sampling distribution of F for the sample size in the "Smiles and Leniency" study. As you can see, it has a positive skew.

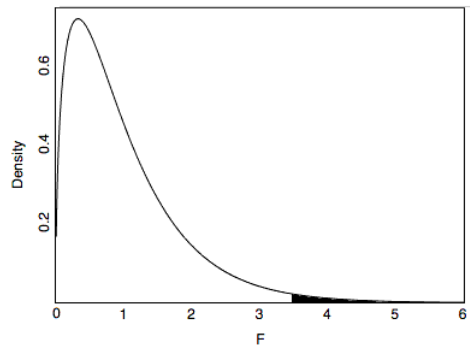


Figure 13.3.1: Distribution of F

From Figure $13.3.1$, you can see that F ratios of 3.465 or above are unusual occurrences. The area to the right of 3.465 represents the probability of an F that large or larger and is equal to 0.018. In other words, given the null hypothesis that all the population means are equal, the probability value is 0.018 and therefore the null hypothesis can be rejected. The conclusion that at least one of the population means is different from at least one of the others is justified.

The shape of the F distribution depends on the sample size. More precisely, it depends on two degrees of freedom (df) parameters: one for the numerator (MSB) and one for the denominator (MSE). Recall that the degrees of freedom for an estimate of variance is equal to the number of observations minus one. Since the MSB is the variance of k means, it has $k-1$ df. The MSE is an average of k variances, each with $n-1$ df. Therefore, the df for MSE is $k(n-1) = N-k$, where N is the total number of observations, n is the number of observations in each group, and k is the number of groups. To summarize:

$$df_{numerator} = k-1 \tag{13.3.6}$$

$$df_{denominator} = N-k \tag{13.3.7}$$

For the "Smiles and Leniency" data,

$$df_{numerator} = k-1 = 4-1 = 3 \tag{13.3.8}$$

$$df_{denominator} = N-k = 136-4 = 132 \tag{13.3.9}$$

$F = 3.465$

The F distribution calculator shows that $p = 0.018$.

F Calculator

One-Tailed or Two?

Is the probability value from an F ratio a one-tailed or a two-tailed probability? In the literal sense, it is a one-tailed probability since, as you can see in Figure $13.3.1$, the probability is the area in the right-hand tail of the distribution. However, the F ratio is sensitive to any pattern of differences among means. It is, therefore, a test of a two-tailed hypothesis and is best considered a two-tailed test.

Relationship to the t test

Since an ANOVA and an independent-groups t test can both test the difference between two means, you might be wondering which one to use. Fortunately, it does not matter since the results will always be the same. When there are only two groups, the following

relationship between F and t will always hold:

$$F(1, dfd) = t^2(df) \tag{13.3.10}$$

where dfd is the degrees of freedom for the denominator of the F test and df is the degrees of freedom for the t test. dfd will always equal df.

Sources of Variation

Why do scores in an experiment differ from one another? Consider the scores of two subjects in the "Smiles and Leniency" study: one from the "False Smile" condition and one from the "Felt Smile" condition. An obvious possible reason that the scores could differ is that the subjects were treated differently (they were in different conditions and saw different stimuli). A second reason is that the two subjects may have differed with regard to their tendency to judge people leniently. A third is that, perhaps, one of the subjects was in a bad mood after receiving a low grade on a test. You can imagine that there are innumerable other reasons why the scores of the two subjects could differ. All of these reasons except the first (subjects were treated differently) are possibilities that were not under experimental investigation and, therefore, all of the differences (variation) due to these possibilities are unexplained. It is traditional to call unexplained variance error even though there is no implication that an error was made. Therefore, the variation in this experiment can be thought of as being either variation due to the condition the subject was in or due to error (the sum total of all reasons the subjects' scores could differ that were not measured).

One of the important characteristics of ANOVA is that it partitions the variation into its various sources. In ANOVA, the term sum of squares (SSQ) is used to indicate variation. The total variation is defined as the sum of squared differences between each score and the mean of all subjects. The mean of all subjects is called the grand mean and is designated as GM. (When there is an equal number of subjects in each condition, the grand mean is the mean of the condition means.) The total sum of squares is defined as

$$SSQ_{total} = \sum (X - GM)^2 \tag{13.3.11}$$

which means to take each score, subtract the grand mean from it, square the difference, and then sum up these squared values. For the "Smiles and Leniency" study, $SSQ_{total} = 377.19$.

The sum of squares condition is calculated as shown below.

$$SSQ_{condition} = n\left[(M_1 - GM)^2 + (M_2 - GM)^2 + \cdots + (M_k - GM)^2\right] \tag{13.3.12}$$

where n is the number of scores in each group, k is the number of groups, M_1 is the mean for Condition 1, M_2 is the mean for Condition 2, and M_k is the mean for Condition k. For the Smiles and Leniency study, the values are:

$$\begin{aligned} SSQ_{condition} &= 34\left[(5.37 - 4.83)^2 + (4.91 - 4.83)^2 + (4.91 - 4.83)^2 + (4.12 - 4.83)^2\right] \\ &= 27.5 \end{aligned}$$

If there are unequal sample sizes, the only change is that the following formula is used for the sum of squares condition:

$$SSQ_{condition} = n_1(M_1 - GM)^2 + n_2(M_2 - GM)^2 + \cdots + n_k(M_k - GM)^2 \tag{13.3.13}$$

where n_i is the sample size of the ith condition. SSQ_{total} is computed the same way as shown above.

The sum of squares error is the sum of the squared deviations of each score from its group mean. This can be written as

$$SSQ_{error} = \sum (X_{i1} - M_1)^2 + \sum (X_{i2} - M_2)^2 + \cdots + \sum (X_{ik} - M_k)^2 \tag{13.3.14}$$

where X_{i1} is the ith score in group 1 and M_1 is the mean for group 1, X_{i2} is the ith score in group 2 and M_2 is the mean for group 2, etc. For the "Smiles and Leniency" study, the means are: 5.368, 4.912, 4.912, and 4.118. The SSQ_{error} is therefore:

$$\begin{aligned} SSQ_{error} &= (2.5 - 5.368)^2 + (5.5 - 5.368)^2 + \ldots + (6.5 - 4.118)^2 \\ &= 349.65 \end{aligned}$$

The sum of squares error can also be computed by subtraction:

$$SSQ_{error} = SSQ_{total} - SSQ_{condition} \tag{13.3.15}$$

$$\begin{aligned} SSQ_{error} &= 377.189 - 27.535 \\ &= 349.65 \end{aligned}$$

Therefore, the total sum of squares of 377.19 can be partitioned into $SSQ_{condition}(27.53)$ and $SSQ_{error}(349.66)$.

Once the sums of squares have been computed, the mean squares (MSB and MSE) can be computed easily. The formulas are:

$$MSB = \frac{SSQ_{condition}}{dfn} \tag{13.3.16}$$

where dfn is the degrees of freedom numerator and is equal to $k-1=3$.

$$MSB = \frac{27.535}{3} = 9.18 \tag{13.3.17}$$

which is the same value of MSB obtained previously (except for rounding error). Similarly,

$$MSE = \frac{SSQ_{error}}{dfd} \tag{13.3.18}$$

where dfd is the degrees of freedom for the denominator and is equal to $N-k$.

$dfd = 136 - 4 = 132$

$MSE = 349.66/132 = 2.65$

which is the same as obtained previously (except for rounding error). Note that the dfd is often called the dfe for degrees of freedom error.

The Analysis of Variance Summary Table shown below is a convenient way to summarize the partitioning of the variance. The rounding errors have been corrected.

Table 13.3.1: ANOVA Summary Table

Source	df	SSQ	MS	F	p
Condition	3	27.5349	9.1783	3.465	0.0182
Error	132	349.6544	2.6489		
Total	135	377.1893			

The first column shows the sources of variation, the second column shows the degrees of freedom, the third shows the sums of squares, the fourth shows the mean squares, the fifth shows the F ratio, and the last shows the probability value. Note that the mean squares are always the sums of squares divided by degrees of freedom. The F and p are relevant only to Condition. Although the mean square total could be computed by dividing the sum of squares by the degrees of freedom, it is generally not of much interest and is omitted here.

Formatting Data for Computer Analysis

Most computer programs that compute ANOVAs require your data to be in a specific form. Consider the data in Table $13.3.3$

Table $13.3.3$: Example Data

Group 1	Group 2	Group 3
3	2	8
4	4	5
5	6	5

Here there are three groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table $13.3.3$is shown in Table $13.3.4$

Table 13.3.4: Reformatted Data

G	Y

1	3
1	4
1	5
2	2
2	4
2	6
3	8
3	5
3	5

To use Analysis Lab to do the calculations, you would copy the data and then

1. Click the "Enter/Edit Data" button. (You may be warned that for security reasons you must use the keyboard shortcut for pasting data.)
2. Paste your data.
3. Click "Accept Data."
4. Set the Dependent Variable to Y.
5. Set the Grouping Variable to G.
6. Click the ANOVA button.

You will find that $F = 1.5$ and $p = 0.296$.

13.4: One-Way Demo

Learning Objectives

- State how the sums of squares are divided among sources of variation
- State the components of an ANOVA summary table and how they relate to each other

Instructions

This simulation demonstrates the partitioning of sums of squares in analysis of variance (ANOVA). Initially, you see a display of the first sample dataset. There are three groups of subjects and four subjects per group. Each score is represented by a small black rectangle; the mean is represented as a red horizontal line. The values range from 0 to 10. The Y-axis ticks represent values of 0, 2.25, 5, 7.75, and 10. The means of the three groups are 2.50, 5.50, and 8.50.

The table in the lower left-hand portion of the window displays, for each group, the sample size, the mean, the sum of squared deviations of individual scores from the mean, and the squared difference between the group mean and the grand mean multiplied by the sample size. The bottom row shows the sums of these quantities. The sum of the last column is the sums of squares between while the sum of the second-to-last column is the sum of squares within. The ANOVA summary table based on these values is shown to the right. The sum of squares between and within are depicted graphically above the ANOVA summary table.

You can choose other datasets using the pop-up menu. You can also enter your own data by clicking on the data display. If you click in a blank area, an new data point is created (if you hold the mouse down after you click you can position the point by moving the mouse). You can modify the data by clicking on a point and moving it to another location. Finally, you can delete a data point by dragging it outside the colored region.

1. Notice how the sum of squares total is divided up into the sum of squared differences of each score from its group mean and the sum of squared differences of the group mean from the grand mean (GM: the mean of all data). Keep in mind that differences of group means from the grand mean have to be multiplied by the sample size.
2. Add a few data points by clicking and note the effects on the sums of squares. Notice that if the points are far from the group mean then the sum of squares within increases greatly.
3. Choose $\text{dataset } 2$. Notice that the means are very different and the data points are all near their group mean. This results in a large sum of squares between and a msall sum of squares within.
4. Look at $\text{dataset } 4$ which is similar but has more subjects.
5. Look at $\text{dataset } 6$ for which the group means are all the same. Note the value of the sum of squares between.
6. Choose "blank dataset" and enter your own data.

Illustrated Instructions

Video Demo

The video demonstration increases the mean of $\text{group } 1$ by dragging the individual points in the group. Notice how the statistics update as the points are moved.

13.5: Multi-Factor Between-Subjects

Learning Objectives

- Define main effect, simple effect, interaction, and marginal mean
- State the relationship between simple effects and interaction
- Compute the source of variation and df for each effect in a factorial design
- Plot the means for an interaction
- Define three-way interaction

Basic Concepts and Terms

In the "Bias Against Associates of the Obese" case study, the researchers were interested in whether the weight of a companion of a job applicant would affect judgments of a male applicant's qualifications for a job. Two independent variables were investigated:

1. whether the companion was obese or of typical weight and
2. whether the companion was a girlfriend or just an acquaintance.

One approach could have been to conduct two separate studies, one with each independent variable. However, it is more efficient to conduct one study that includes both independent variables. Moreover, there is a much bigger advantage than efficiency for including two variables in the same study: it allows a test of the interaction between the variables. There is an interaction when the effect of one variable differs depending on the level of a second variable. For example, it is possible that the effect of having an obese companion would differ depending on the relationship to the companion. Perhaps there is more prejudice against a person with an obese companion if the companion is a girlfriend than if she is just an acquaintance. If so, there would be an interaction between the obesity factor and the relationship factor.

There are three effects of interest in this experiment:

1. Weight: Are applicants judged differently depending on the weight of their companion?
2. Relationship: Are applicants judged differently depending on their relationship with their companion?
3. Weight x Relationship Interaction: Does the effect of weight differ depending on the relationship with the companion?

The first two effects (Weight and Relationship) are both main effects. A main effect of an independent variable is the effect of the variable averaging over the levels of the other variable(s). It is convenient to talk about main effects in terms of marginal means. A marginal mean for a level of a variable is the mean of the means of all levels of the other variable. For example, the marginal mean for the level "Obese" is the mean of "Girlfriend Obese" and "Acquaintance Obese." Table $13.5.1$ shows that this marginal mean is equal to the mean of 5.65 and 6.15, which is 5.90. Similarly, the marginal mean for the level "Typical" is the mean of 6.19 and 6.59, which is 6.39. The main effect of Weight is based on a comparison of these two marginal means. Similarly, the marginal means for "Girlfriend" and "Acquaintance" are 5.92 and 6.37.

Table 13.5.1: Means for All Four Conditions

	Companion Weight			
Obese	Typical	Marginal Mean		
Relationship	Girlfriend	5.65	6.19	5.92
	Acquaintance	6.15	6.59	6.37
	Marginal Mean	5.90	6.39	

In contrast to a main effect, which is the effect of a variable averaged across levels of another variable, the simple effect of a variable is the effect of the variable at a single level of another variable. The simple effect of Weight at the level of "Girlfriend" is the difference between the "Girlfriend Typical" and the "Girlfriend Obese" conditions. The difference is $6.19 - 5.65 = 0.54$. Similarly, the simple effect of Weight at the level of "Acquaintance" is the difference between the "Acquaintance Typical" and the "Acquaintance Obese" conditions. The difference is $6.59 - 6.15 = 0.44$.

Recall that there is an interaction when the effect of one variable differs depending on the level of another variable. This is equivalent to saying that there is an interaction when the simple effects differ. In this example, the simple effects of weight are 0.54

and 0.44. As shown below, these simple effects are not significantly different.

Tests of Significance

The important questions are not whether there are main effects and interactions in the sample data. Instead, what is important is what the sample data allow you to conclude about the population. This is where Analysis of Variance comes in. ANOVA tests main effects and interactions for significance. An ANOVA Summary Table for these data is shown in Table $13.5.2$

Table 13.5.2: ANOVA Summary Table

Source	df	SSQ	MS	F	p
Weight	1	10.4673	10.4673	6.214	0.0136
Relation	1	8.8144	8.8144	5.233	0.0234
W x R	1	0.1038	0.1038	0.062	0.8043
Error	172	289.7132	1.6844		
Total	175	310.1818			

Consider first the effect of "Weight." The degrees of freedom (df) for "Weight" is 1. The degrees of freedom for a main effect is always equal to the number of levels of the variable minus one. Since there are two levels of the "Weight" variable (typical and obese), the df is $2-1=1$. We skip the calculation of the sum of squares (SSQ) not because it is difficult, but because it is so much easier to rely on computer programs to compute it. The mean square (MS) is the sum of squares divided by the df. The F ratio is computed by dividing the MS for the effect by the MS for error (MSE). For the effect of "Weight," $F=10.4673/1.6844=6.214$The last column, p, is the probability of getting an F of 6.214or larger given that there is no effect of weight in the population. The p value is $0.0136$and therefore the null hypothesis of no main effect of "Weight" is rejected. The conclusion is that being accompanied by an obese companion lowers judgments of qualifications.

The effect "Relation" is interpreted the same way. The conclusion is that being accompanied by a girlfriend leads to lower ratings than being accompanied by an acquaintance.

The df for an interaction is the product of the $df's$ of variables in the interaction. For the "Weight x Relation" interaction ($W \times R$), the $df=1$ since both Weight and Relation have one $df: 1 \times 1=1$. The p value for the interaction is 0.8043, which is the probability of getting an interaction as big or bigger than the one obtained in the experiment if there were no interaction in the population. Therefore, these data provide no evidence for an interaction. Always keep in mind that the lack of evidence for an effect does not justify the conclusion that there is no effect. In other words, you do not accept the null hypothesis just because you do not reject it.

For "Error," the degrees of freedom is equal to the total number of observations minus the total number of groups. The sample sizes of the four conditions in this experiment are shown in Table $13.5.3$. The total number of observations is $40+42+40+54=176$. Since there are four groups, $dfe=176-4=172$.

Table 13.5.3: Sample Sizes for All Four Conditions

		Companion Weight	
Obese	Typical		
Relationship	Girlfriend	40	42
	Acquaintance	40	54

The final row in the ANOVA Summary Table is "Total." The degrees of freedom total is equal to the sum of all degrees of freedom. It is also equal to the $number\ of\ observations\ minus\ 1$, or $176-1=175$. When there are equal sample sizes, the sum of squares total will equal the sum of all other sums of squares. However, when there are unequal sample sizes, as there are here, this will not generally be true. The reasons for this are complex and are discussed in the section Unequal Sample Sizes.

Plotting Means

Although the plot shown in Figure 13.5.1 illustrates the main effects as well as the interaction (or lack of an interaction), it is called an interaction plot. It is important to consider the components of this plot carefully. First, the dependent variable is on the Y-axis. Second, one of the independent variables is on the X-axis. In this case, it is the variable "Weight." Finally, a separate line is drawn for each level of the other independent variable. It is better to label the lines right on the graph, as shown here, than with a legend.

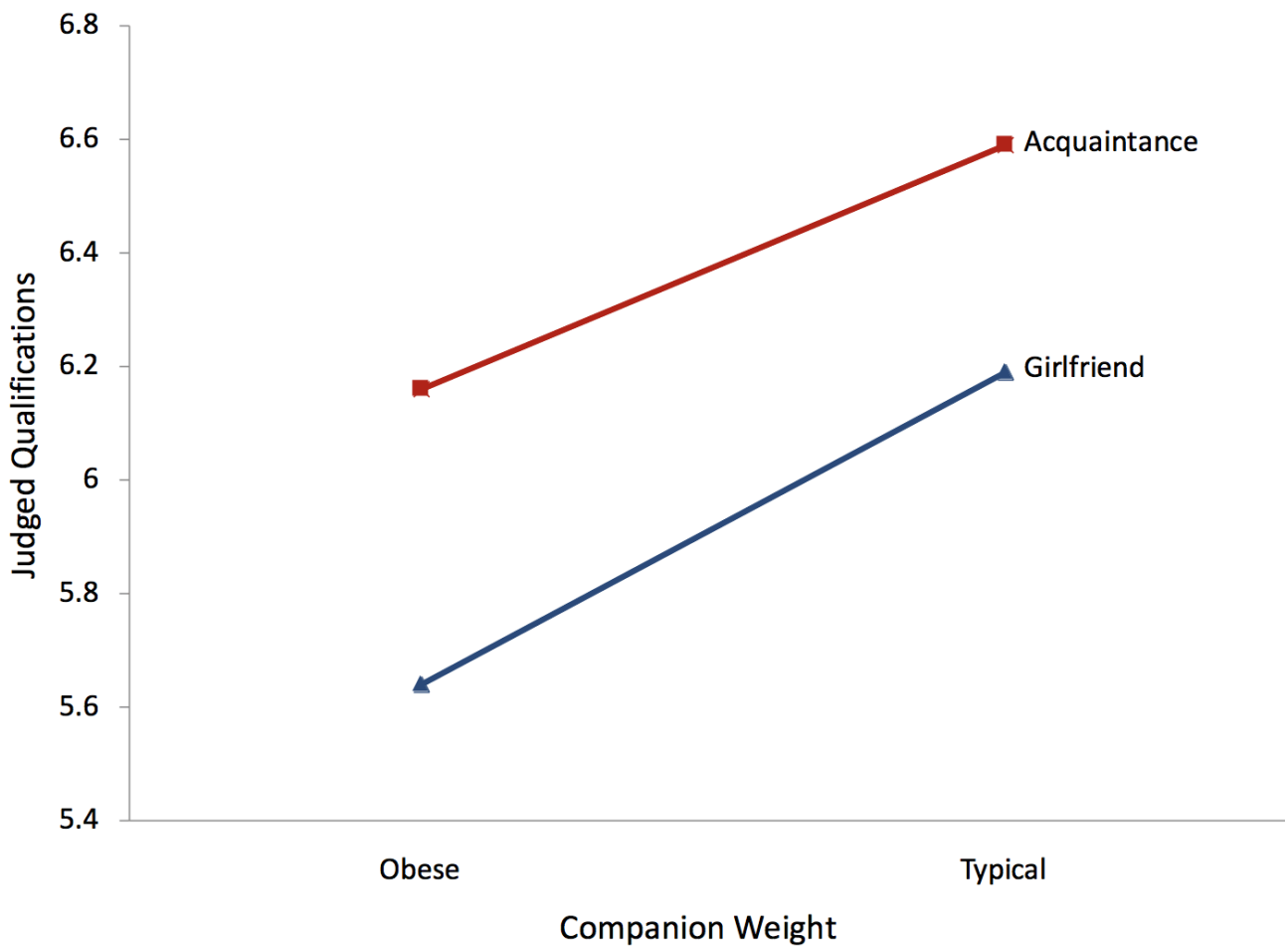


Figure 13.5.1: An Interaction Plot

If you have three or more levels on the X-axis, you should not use lines unless there is some numeric ordering to the levels. If your variable on the X-axis is a qualitative variable, you can use a plot such as the one in Figure 13.5.2 However, as discussed in the section on bar charts, it would be better to replace each bar with a box plot.

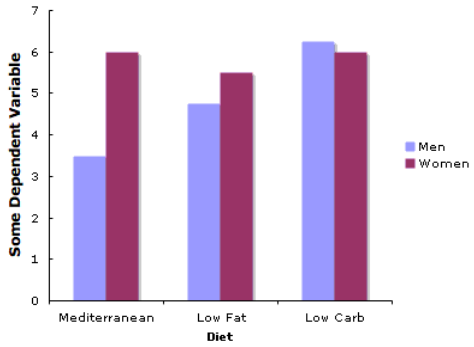


Figure 13.5.2: Plot With a Qualitative Variable on the X-axis

Figure 13.5.3 shows such a plot. Notice how it contains information about the medians, quantiles, and minimums and maximums not contained in Figure 13.5.2 Most important, you get an idea about how much the distributions overlap from Figure 13.5.3 which you do not get from Figure 13.5.2

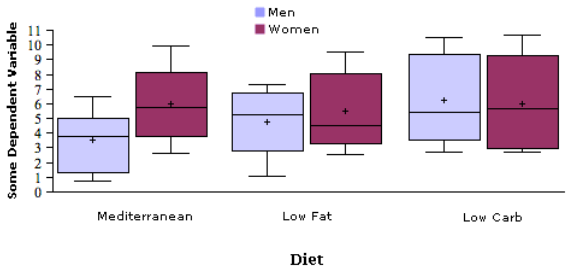


Figure **13.5.3:** Box plots

Line graphs are a good option when there are more than two levels of a numeric variable. Figure 13.5.4 shows an example. A line graph has the advantage of showing the pattern of interaction clearly. Its disadvantage is that it does not convey the distributional information contained in box plots.

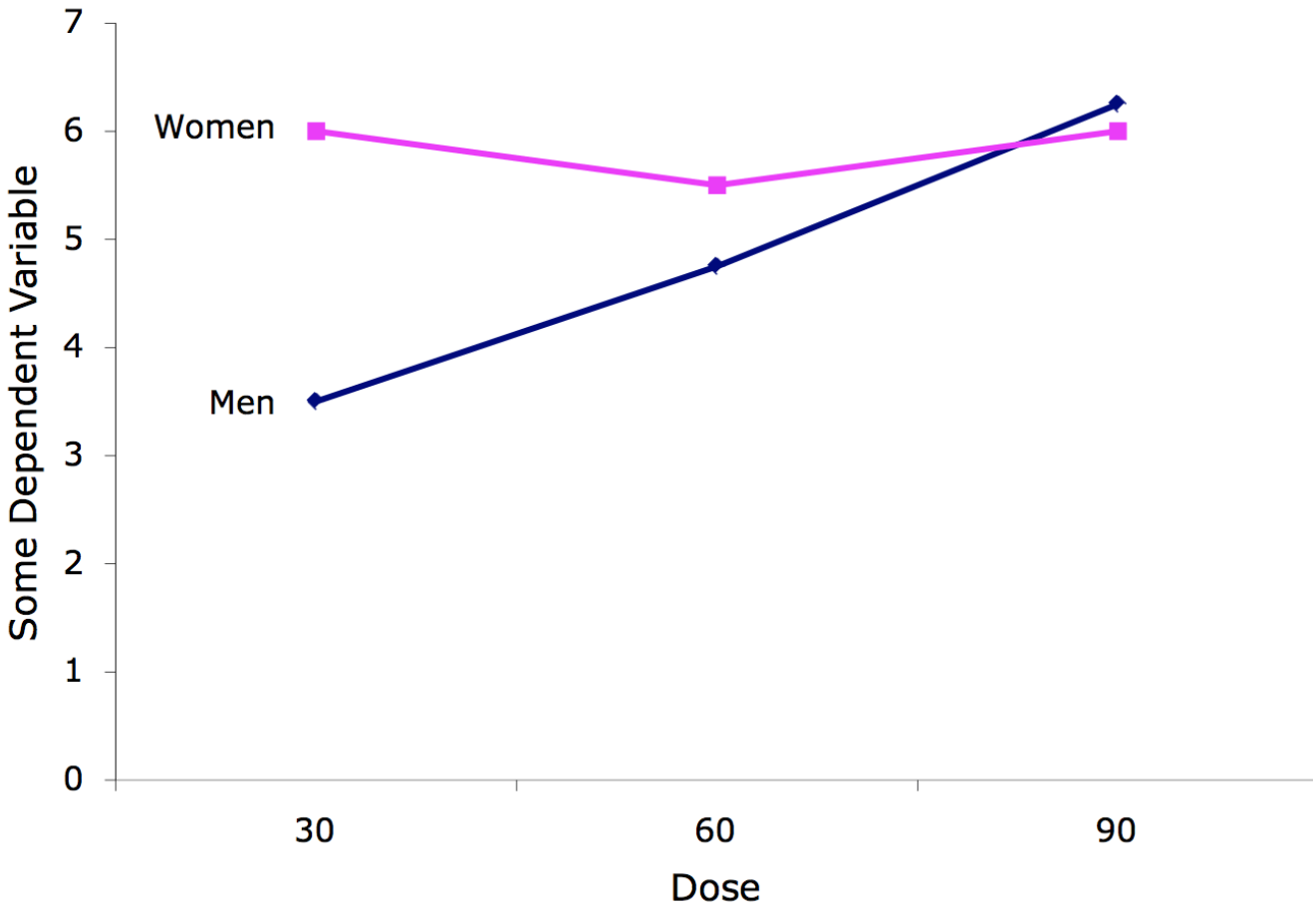


Figure 13.5.1: Plot With a Quantitative Variable on the X-axis

An Example with Interaction

The following example was presented in the section on specific comparisons among means. It is also relevant here.

This example uses the made-up data from a hypothetical experiment shown in Table 13.5.4 Twelve subjects were selected from a population of high-self-esteem subjects and an additional 12 subjects were selected from a population of low-self-esteem subjects. Subjects then performed on a task and (independent of how well they really did) half in each esteem category were told they

succeeded and the other half were told they failed. Therefore, there were six subjects in each of the four esteem/outcome combinations and 24 subjects in all.

After the task, subjects were asked to rate (on a 10-point scale) how much of their outcome (success or failure) they attributed to themselves as opposed to being due to the nature of the task.

Table 13.5.4: Data from Hypothetical Experiment on Attribution

		Esteem	
High	Low		
Outcome	Success	7	6
		8	5
		7	7
		8	4
		9	5
		5	6
	Failure	4	9
		6	8
		5	9
		4	8
		7	7
		3	6

The ANOVA Summary Table for these data is shown in Table 13.5.5

Table 13.5.5: ANOVA Summary Table for Made-Up Data

Source	df	SSQ	MS	F	p
Outcome	1	0.0417	0.0417	0.0256	0.8744
Esteem	1	2.0417	2.0417	1.2564	0.2756
O x E	1	35.0417	35.0417	21.5641	0.0002
Error	20	32.5000	1.6250		
Total	23	69.6250			

As you can see, the only significant effect is the $Outcome \times Esteem$ $(O \times E)$ interaction. The form of the interaction can be seen in Figure 13.5.5

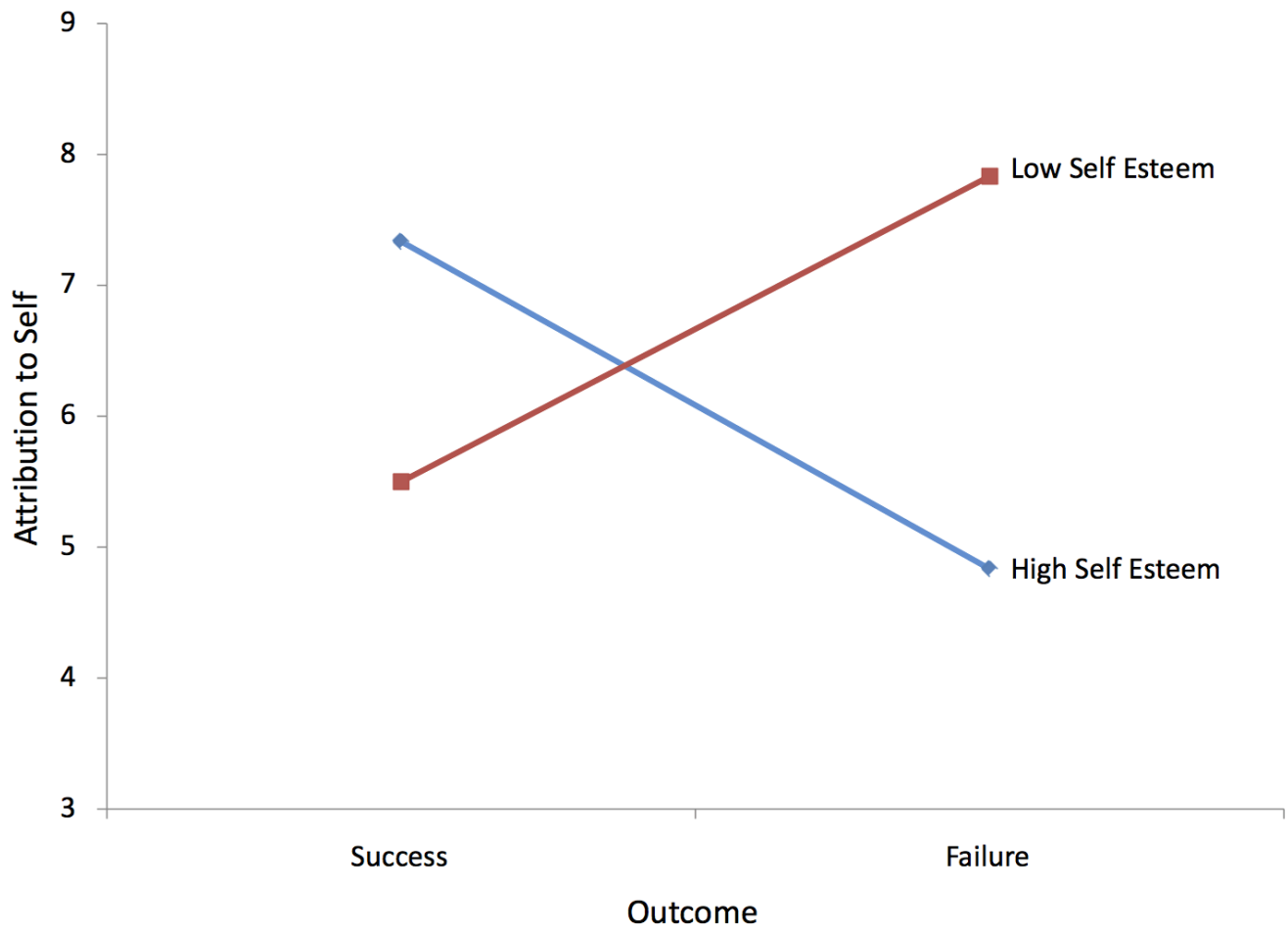


Figure 13.5.5: Interaction Plot for Made-Up Data

Clearly the effect of "Outcome" is different for the two levels of "Esteem": For subjects high in self-esteem, failure led to less attribution to oneself than did success. By contrast, for subjects low in self-esteem, failure led to more attribution to oneself than did success. Notice that the two lines in the graph are not parallel. Nonparallel lines indicate interaction. The significance test for the interaction determines whether it is justified to conclude that the lines in the population are not parallel. Lines do not have to cross for there to be an interaction.

Three-Factor Designs

Three-factor designs are analyzed in much the same way as two-factor designs. Table 13.5.6 shows the analysis of a study described by Franklin and Cooley investigating three factors on the strength of industrial fans:

1. Hole Shape (Hex or Round)
2. Assembly Method (Staked or Spun), and
3. Barrel Surface (Knurled or Smooth).

The dependent variable, Breaking Torque, was measured in foot-pounds. There were eight observations in each of the eight combinations of the three factors.

As you can see in Table 13.5.6, there are three main effects, three two-way interactions, and one three-way interaction. The degrees of freedom for the main effects are, as in a two-factor design, equal to the number of levels of the factor minus one. Since all the factors here have two levels, all the main effects have one degree of freedom. The interaction degrees of freedom is always equal to the product of the degrees of freedom of the component parts. This holds for the three-factor interaction as well as for the two-factor interactions. The error degrees of freedom is equal to the number of observations (64) minus the number of groups (8) and equals 56.

Table 13.5.6: ANOVA Summary Table for Fan Data

Source	df	SSQ	MS	F	p
Hole	1	8258.27	8258.27	266.68	<0.0001
Assembly	1	13369.14	13369.14	431.73	<0.0001
H x A	1	2848.89	2848.89	92.00	<0.0001
Barrel	1	35.0417	35.0417	21.5641	<0.0001
H x B	1	594.14	594.14	19.1865	<0.0001
A x B	1	135.14	135.14	4.36	0.0413
H x A x B	1	1396.89	1396.89	45.11	<0.0001
Error	56	1734.12	30.97		
Total	63	221386.91			

A three-way interaction means that the two-way interactions differ as a function of the level of the third variable. The usual way to portray a three-way interaction is to plot the two-way interactions separately. Figure 13.5.6 shows the Barrel (Knurled or Smooth) × Assembly (Staked or Spun) separately for the two levels of Hole Shape (Hex or Round). For the Hex Shape, there is very little interaction with the lines being close to parallel with a very slight tendency for the effect of Barrel to be bigger for Staked than for Spun. The two-way interaction for the Round Shape is different: The effect of Barrel is bigger for Spun than for Staked. The finding of a significant three-way interaction indicates that this difference in two-way interactions is significant.

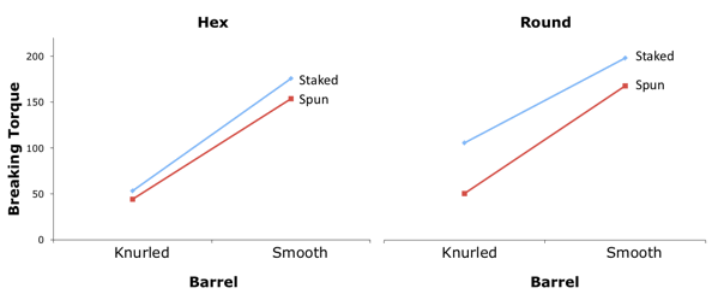


Figure 13.5.6: Plot of the three-way interaction

Formatting Data for Computer Analysis

The data in Table 13.5.4 have been reformatted in Table 13.5.7. Note how there is one column to indicate the level of outcome and one column to indicate the level of esteem. The coding is as follows:

High self-esteem: 1
Low self-esteem: 2

Success: 1
Failure: 2

Table 13.5.7: Attribution Data Reformatted

outcome	esteem	attrib
1	1	7
1	1	8
1	1	7
1	1	8
1	1	9

outcome	esteem	attrib
1	1	5
1	2	6
1	2	5
1	2	7
1	2	4
1	2	5
1	2	6
2	1	4
2	1	6
2	1	5
2	1	4
2	1	7
2	1	3
2	2	9
2	2	8
2	2	9
2	2	8
2	2	7
2	2	6

To use Analysis Lab to do the calculations, you would copy the data and then

1. Click the "Enter/Edit Data" button. (You may be warned that for security reasons you must use the keyboard shortcut for pasting data.)
2. Paste your data.
3. Click "Accept Data."
4. Click the "Advanced" button next to the "ANOVA" button.
5. Select "attrib" as the dependent variable and both "outcome" and "esteem" as "group" variables.
6. Click the "Do ANOVA" button.

13.6: Unequal Sample Sizes

Learning Objectives

- State why unequal n can be a problem
- Define confounding
- Compute weighted and unweighted means
- Distinguish between **Type I** and **Type III** sums of squares
- Describe why the cause of the unequal sample sizes makes a difference in the interpretation

The Problem of Confounding

Whether by design, accident, or necessity, the number of subjects in each of the conditions in an experiment may not be equal. For example, the sample sizes for the "Bias Against Associates of the Obese" case study are shown in Table $13.6.1$. Although the sample sizes were approximately equal, the "Acquaintance Typical" condition had the most subjects. Since n is used to refer to the sample size of an individual group, designs with unequal sample sizes are sometimes referred to as designs with unequal n.

Table 13.6.1: Sample Sizes for "Bias Against Associates of the Obese" Study

		Companion Weight	
Obese	Typical		
Relationship	Girlfriend	40	42
	Acquaintance	40	54

We consider an absurd design to illustrate the main problem caused by unequal n. Suppose an experimenter were interested in the effects of diet and exercise on cholesterol. The sample sizes are shown in Table $13.6.2$

Table 13.6.2: Sample Sizes for "Diet and Exercise" Example

		Exercise	
Moderate	None		
Diet	Low Fat	5	0
	High Fat	0	5

What makes this example absurd is that there are no subjects in either the "Low-Fat No-Exercise" condition or the "High-Fat Moderate-Exercise" condition. The hypothetical data showing change in cholesterol are shown in Table $13.6.3$.

Table 13.6.3: Data for "Diet and Exercise" Example

		Exercise		
Moderate	None	Mean		
Diet	Low Fat	-20		-25
		-25		
		-30		
		-35		
		-15		
	High Fat		-20	-5
			6	
			-10	
			-6	

			Exercise	
			5	
	Mean	-25	-5	-15

The last column shows the mean change in cholesterol for the two Diet conditions, whereas the last row shows the mean change in cholesterol for the two Exercise conditions. The value of -15 in the lower-right-most cell in the table is the mean of all subjects.

We see from the last column that those on the low-fat diet lowered their cholesterol an average of 25 units, whereas those on the high-fat diet lowered theirs by only an average of 5 units. However, there is no way of knowing whether the difference is due to diet or to exercise since every subject in the low-fat condition was in the moderate-exercise condition and every subject in the high-fat condition was in the no-exercise condition. Therefore, Diet and Exercise are completely confounded. The problem with unequal n is that it causes confounding.

Weighted and Unweighted Means

The difference between weighted and unweighted means is a difference critical for understanding how to deal with the confounding resulting from unequal n.

Weighted and unweighted means will be explained using the data shown in Table $13.6.4$ Here, Diet and Exercise are confounded because 80% of the subjects in the low-fat condition exercised as compared to 20% of those in the high-fat condition. However, there is not complete confounding as there was with the data in Table $13.6.3$

The weighted mean for "Low Fat" is computed as the mean of the "Low-Fat Moderate-Exercise" mean and the "Low-Fat No-Exercise" mean, weighted in accordance with sample size. To compute a weighted mean, you multiply each mean by its sample size and divide by N, the total number of observations. Since there are four subjects in the "Low-Fat Moderate-Exercise" condition and one subject in the "Low-Fat No-Exercise" condition, the means are weighted by factors of 4 and 1 as shown below, where M_W is the weighted mean.

$$M_W = \frac{(4)(-27.5)+(1)(-20)}{5} = -26 \tag{13.6.1}$$

The weighted mean for the low-fat condition is also the mean of all five scores in this condition. Thus if you ignore the factor "Exercise," you are implicitly computing weighted means.

The unweighted mean for the low-fat condition (M_U) is simply the mean of the two means.

$$M_U = \frac{-27.5-20}{2} = -23.75 \tag{13.6.2}$$

Table $13.6.4$: Data for Diet and Exercise with Partial Confounding Example

		Exercise			
Moderate	None	Weighted Mean	Unweighted Mean		
Diet	Low Fat	-20	-20	-26	-23.750
		-25			
		-30			
		-35			
		M=-27.5	M=-20.0		
	High Fat	-15	6	-4	-8.125
			-6		
			5		
			-10		
		M=-15.0	M=-1.25		

		Exercise		
	Weighted Mean	-25	-5	
	Unweighted Mean	-21.25	-10.625	

One way to evaluate the main effect of Diet is to compare the weighted mean for the low-fat diet (-26) with the weighted mean for the high-fat diet (-4). This difference of -22 is called "the effect of diet ignoring exercise" and is misleading since most of the low-fat subjects exercised and most of the high-fat subjects did not. However, the difference between the unweighted means of $-15.625((-23.750)-(-8.125))$ is not affected by this confounding and is therefore a better measure of the main effect. In short, weighted means ignore the effects of other variables (exercise in this example) and result in confounding; unweighted means control for the effect of other variables and therefore eliminate the confounding.

Statistical analysis programs use different terms for means that are computed controlling for other effects. SPSS calls them estimated marginal means, whereas SAS and SAS JMP call them least squares means.

Types of Sums of Squares

The section on Multi-Factor ANOVA stated that when there are unequal sample sizes, the sum of squares total is not equal to the sum of the sums of squares for all the other sources of variation. This is because the confounded sums of squares are not apportioned to any source of variation. For the data in Table $13.6.4$, the sum of squares for Diet is 390.625, the sum of squares for Exercise is 180.625, and the sum of squares confounded between these two factors is 819.375 (the calculation of this value is beyond the scope of this introductory text). In the ANOVA Summary Table shown in Table $13.6.5$, this large portion of the sums of squares is not apportioned to any source of variation and represents the "missing" sums of squares. That is, if you add up the sums of squares for Diet, Exercise, $D \times E$, and Error, you get 902.625. If you add the confounded sum of squares of 819.375 to this value, you get the total sum of squares of 1722.000. When confounded sums of squares are not apportioned to any source of variation, the sums of squares are called **Type III** sums of squares. **Type III** sums of squares are, by far, the most common and if sums of squares are not otherwise labeled, it can safely be assumed that they are **Type III**.

Table $13.6.5$: ANOVA Summary Table for Type III SSQ

Source	df	SSQ	MS	F	p
Diet	1	390.625	390.625	7.42	0.034
Exercise	1	180.625	180.625	3.43	0.113
D x E	1	15.625	15.625	0.30	0.605
Error	6	315.750	52.625		
Total	9	1722.000			

When all confounded sums of squares are apportioned to sources of variation, the sums of squares are called **Type I** sums of squares. The order in which the confounded sums of squares are apportioned is determined by the order in which the effects are listed. The first effect gets any sums of squares confounded between it and any of the other effects. The second gets the sums of squares confounded between it and subsequent effects, but not confounded with the first effect, etc. The **Type I** sums of squares are shown in Table $13.6.6$. As you can see, with **Type I** sums of squares, the sum of all sums of squares is the total sum of squares.

Table 13.6.6: ANOVA Summary Table for Type I SSQ

Source	df	SSQ	MS	F	p
Diet	1	1210.000	1210.000	22.99	0.003
Exercise	1	180.625	180.625	3.43	0.113
D x E	1	15.625	15.625	0.30	0.605
Error	6	315.750	52.625		
Total	9	1722.000			

In **Type II** sums of squares, sums of squares confounded between main effects are not apportioned to any source of variation, whereas sums of squares confounded between main effects and interactions are apportioned to the main effects. In our example, there is no confounding between the $D \times E$ interaction and either of the main effects. Therefore, the **Type II** sums of squares are equal to the Type III sums of squares.

Which Type of Sums of Squares to Use (optional)

Type I sums of squares allow the variance confounded between two main effects to be apportioned to one of the main effects. Unless there is a strong argument for how the confounded variance should be apportioned (which is rarely, if ever, the case), **Type I** sums of squares are not recommended.

There is not a consensus about whether **Type II** or **Type III** sums of squares is to be preferred. On the one hand, if there is no interaction, then **Type II** sums of squares will be more powerful for two reasons:

1. variance confounded between the main effect and interaction is properly assigned to the main effect and
2. weighting the means by sample sizes gives better estimates of the effects.

To take advantage of the greater power of **Type II** sums of squares, some have suggested that if the interaction is not significant, then **Type II** sums of squares should be used. Maxwell and Delaney (2003) caution that such an approach could result in a **Type II** error in the test of the interaction. That is, it could lead to the conclusion that there is no interaction in the population when there really is one. This, in turn, would increase the **Type I** error rate for the test of the main effect. As a result, their general recommendation is to use **Type III** sums of squares.

Maxwell and Delaney (2003) recognized that some researchers prefer **Type II** sums of squares when there are strong theoretical reasons to suspect a lack of interaction and the p value is much higher than the typical α level of 0.05. However, this argument for the use of **Type II** sums of squares is not entirely convincing. As Tukey (1991) and others have argued, it is doubtful that any effect, whether a main effect or an interaction, is exactly 0 in the population. Incidentally, Tukey argued that the role of significance testing is to determine whether a confident conclusion can be made about the direction of an effect, not simply to conclude that an effect is not exactly 0.

Finally, if one assumes that there is no interaction, then an ANOVA model with no interaction term should be used rather than **Type II** sums of squares in a model that includes an interaction term. (Models without interaction terms are not covered in this book).

There are situations in which **Type II** sums of squares are justified even if there is strong interaction. This is the case because the hypotheses tested by **Type II** and **Type III** sums of squares are different, and the choice of which to use should be guided by which hypothesis is of interest. Recall that **Type II** sums of squares weight cells based on their sample sizes whereas **Type III** sums of squares weight all cells the same. Consider Figure $13.6.1$ which shows data from a hypothetical $A(2) \times B(2)$ design. The sample sizes are shown numerically and are represented graphically by the areas of the endpoints.

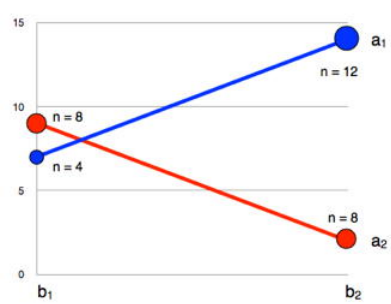


Figure 13.6.1: An interaction plot with unequal sample sizes

First, let's consider the hypothesis for the main effect of B tested by the **Type III** sums of squares. **Type III** sums of squares weight the means equally and, for these data, the marginal means for b_1 and b_2 are equal:

For $b_1 : (b_1a_1 + b_1a_2)/2 = (7+9)/2 = 8$

For $b_2 : (b_2a_1 + b_2a_2)/2 = (14+2)/2 = 8$

Thus, there is no main effect of B when tested using **Type III** sums of squares. For **Type II** sums of squares, the means are weighted by sample size.

For $b_1 : (4 \times b_1a_1 + 8 \times b_1a_2)/12 = (4 \times 7 + 8 \times 9)/12 = 8.33$

For $b_2 : (12 \times b_2 a_1 + 8 \times b_2 a_2)/20 = (12 \times 14 + 8 \times 2)/20 = 9.2$

Since the weighted marginal mean for b_2 is larger than the weighted marginal mean for b_1, there is a main effect of B when tested using **Type II** sums of squares.

The **Type II** and **Type III** analysis are testing different hypotheses. First, let's consider the case in which the differences in sample sizes arise because in the sampling of intact groups, the sample cell sizes reflect the population cell sizes (at least approximately). In this case, it makes sense to weight some means more than others and conclude that there is a main effect of B. This is the result obtained with **Type II** sums of squares. However, if the sample size differences arose from random assignment, and there just happened to be more observations in some cells than others, then one would want to estimate what the main effects would have been with equal sample sizes and, therefore, weight the means equally. With the means weighted equally, there is no main effect of B, the result obtained with **Type III** sums of squares.

Unweighted Means Analysis

Type III sums of squares are tests of differences in unweighted means. However, there is an alternative method to testing the same hypotheses tested using **Type III** sums of squares. This method, unweighted means analysis, is computationally simpler than the standard method but is an approximate test rather than an exact test. It is, however, a very good approximation in all but extreme cases. Moreover, it is exactly the same as the traditional test for effects with one degree of freedom. The Analysis Lab uses unweighted means analysis and therefore may not match the results of other computer programs exactly when there is unequal n and the df are greater than one.

Causes of Unequal Sample Sizes

None of the methods for dealing with unequal sample sizes are valid if the experimental treatment is the source of the unequal sample sizes. Imagine an experiment seeking to determine whether publicly performing an embarrassing act would affect one's anxiety about public speaking. In this imaginary experiment, the experimental group is asked to reveal to a group of people the most embarrassing thing they have ever done. The control group is asked to describe what they had at their last meal. Twenty subjects are recruited for the experiment and randomly divided into two equal groups of 10, one for the experimental treatment and one for the control. Following their descriptions, subjects are given an attitude survey concerning public speaking. This seems like a valid experimental design. However, of the 10 subjects in the experimental group, four withdrew from the experiment because they did not wish to publicly describe an embarrassing situation. None of the subjects in the control group withdrew. Even if the data analysis were to show a significant effect, it would not be valid to conclude that the treatment had an effect because a likely alternative explanation cannot be ruled out; namely, subjects who were willing to describe an embarrassing situation differed from those who were not. Thus, the differential dropout rate destroyed the random assignment of subjects to conditions, a critical feature of the experimental design. No amount of statistical adjustment can compensate for this flaw.

1. Maxwell, S. E., & Delaney, H. D. (2003) *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Second Edition, Lawrence Erlbaum Associates, Mahwah, New Jersey.
2. Tukey, J. W. (1991) The philosophy of multiple comparisons, *Statistical Science*, *6*, 110-116.

13.7: Tests Supplementing

Learning Objectives

- Compute Tukey HSD test
- Describe an interaction in words
- Describe why one might want to compute simple effect tests following a significant interaction

The null hypothesis tested in a one-factor ANOVA is that all the population means are equal. Stated more formally,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \tag{13.7.1}$$

where H_0 is the null hypothesis and k is the number of conditions. When the null hypothesis is rejected, all that can be said is that at least one population mean is different from at least one other population mean. The methods for doing more specific tests described All Pairwise Comparisons among Means and in Specific Comparisons apply here. Keep in mind that these tests are valid whether or not they are preceded by an ANOVA.

Main Effects

As shown below, significant main effects in multi-factor designs can be followed up in the same way as significant effects in one-way designs. Table $13.7.1$ shows the data from an imaginary experiment with three levels of Factor A and two levels of Factor B.

Table 13.7.1: Made-Up Example Data

	A1	A2	A3	Marginal Means
B1	5	9	7	7.08
	4	8	9	
	6	7	9	
	5	8	8	
	Mean = 5	Mean = 8	Mean = 8.25	
B2	4	8	8	6.50
	3	6	9	
	6	8	7	
	8	5	6	
	Mean = 5.25	Mean = 6.75	Mean = 7.50	
Marginal Means	5.125	7.375	7.875	6.79

Table $13.7.2$shows the ANOVA Summary Table for these data. The significant main effect of A indicates that, in the population, at least one of the marginal means for A is different from at least one of the others.

Table 13.7.2: ANOVA Summary Table for Made-Up Example Data

Source	df	SSQ	MS	F	p
A	2	34.333	17.167	9.29	0.0017
B	1	2.042	2.042	1.10	0.3070
A x B	2	2.333	1.167	0.63	0.5431
Error	18	33.250	1.847		
Total	23	71.958			

The Tukey HSD test can be used to test all pairwise comparisons among means in a one-factor ANOVA as well as comparisons among marginal means in a multi-factor ANOVA. The formula for the equal-sample-size case is shown below.

$$Q = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n}}} \tag{13.7.2}$$

where M_i and M_j are marginal means, MSE is the mean square error from the ANOVA, and n is the number of scores each mean is based upon. For this example, $MSE = 1.847$ and $n = 8$ because there are eight scores at each level of A. The probability value can be computed using the Studentized Range Calculator. The degrees of freedom is equal to the degrees of freedom error. For this example, $df = 18$. The results of the Tukey HSD test are shown in Table $13.7.3$. The mean for A_1 is significantly lower than the mean for A_2 and the mean for A_3. The means for A_2 and A_3 are not significantly different.

Table 13.7.3: Pairwise Comparisons Among Marginal Means for A

Comparison	M_i - M_j	Q	p
A_1 - A_2	-2.25	-4.68	0.010
A_1 - A_3	-2.75	-5.72	0.002
A_2 - A_3	-0.50	-1.04	0.746

Specific comparisons among means are also carried out much the same way as shown in the relevant section on testing means. The formula for L is

$$L = \sum c_i M_i \tag{13.7.3}$$

where c_i is the coefficient for the ith marginal mean and M_i is the ith marginal mean. For example, to compare A_1 with the average of A_2 and A_3, the coefficients would be $1, -0.5, -0.5$. Therefore,

$$L = (1)(5.125) + (-0.5)(7.375) + (-0.5)(7.875) = -2.5 \tag{13.7.4}$$

To compute t, use:

$$t = \frac{L}{\sqrt{\frac{\sum c_i^2 MSE}{n}}} = -4.25$$

where MSE is the mean square error from the ANOVA and n is the number of scores each marginal mean is based on (eight in this example). The degrees of freedom is the degrees of freedom error from the ANOVA and is equal to 18. Using the Online Calculator, we find that the two-tailed probability value is 0.0005. Therefore, the difference between A_1 and the average of A_2 and A_3 is significant.

t distribution Calculator

Important issues concerning multiple comparisons and orthogonal comparisons are discussed in the Specific Comparisons section in the Testing Means chapter.

Interactions

The presence of a significant interaction makes the interpretation of the results more complicated. Since an interaction means that the simple effects are different, the main effect as the mean of the simple effects does not tell the whole story. This section discusses how to describe interactions, proper and improper uses of simple effects tests, and how to test components of interactions.

Describing Interactions

A crucial first step in understanding a significant interaction is constructing an interaction plot. Figure $13.7.1$ shows an interaction plot from data presented in the section on Multi-Factor ANOVA.

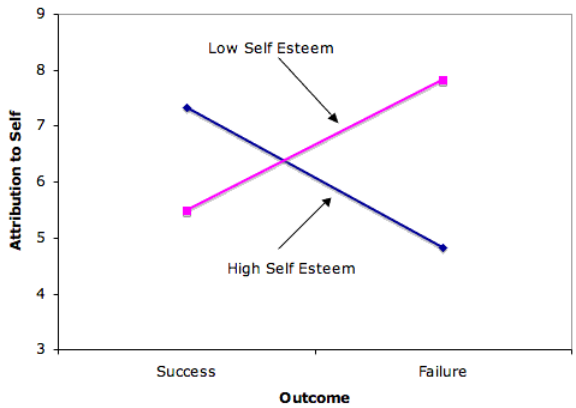


Figure 13.7.1: Interaction Plot for Made-Up Data

The second step is to describe the interaction in a clear and understandable way. This is often done by describing how the simple effects differed. Since this should be done using as little jargon as possible, the expression "simple effect" need not appear in the description. An example is as follows:

The effect of Outcome differed depending on the subject's self-esteem. The difference between the attribution to self following success and the attribution to self following failure was larger for high-self-esteem subjects (mean difference = 2.50) than for low-self-esteem subjects (mean difference = -2.33).

No further analyses are helpful in understanding the interaction since the interaction means only that the simple effects differ. The interaction's significance indicates that the simple effects differ from each other, but provides no information about whether they differ from zero.

Simple Effect Tests

It is not necessary to know whether the simple effects differ from zero in order to understand an interaction because the question of whether simple effects differ from zero has nothing to do with interaction except that if they are both zero there is no interaction. It is not uncommon to see research articles in which the authors report that they analyzed simple effects in order to explain the interaction. However, this is not a valid approach since an interaction does not depend on the analysis of the simple effects.

However, there is a reason to test simple effects following a significant interaction. Since an interaction indicates that simple effects differ, it means that the main effects are not general. In the made-up example, the main effect of Outcome is not very informative, and the effect of outcome should be considered separately for high- and low-self-esteem subjects.

As will be seen, the simple effects of Outcome are significant and in opposite directions: Success significantly increases attribution to self for high-self-esteem subjects and significantly lowers attribution to self for low-self-esteem subjects. This is a very easy result to interpret.

What would the interpretation have been if neither simple effect had been significant? On the surface, this seems impossible: How can the simple effects both be zero if they differ from each other significantly as tested by the interaction? The answer is that a non-significant simple effect does not mean that the simple effect is zero: the null hypothesis should not be accepted just because it is not rejected.

(See section on Interpreting Non-Significant Results)

If neither simple effect is significant, the conclusion should be that the simple effects differ, and that at least one of them is not zero. However, no conclusion should be drawn about which simple effect(s) is/are not zero.

Another error that can be made by mistakenly accepting the null hypothesis is to conclude that two simple effects are different because one is significant and the other is not. Consider the results of an imaginary experiment in which the researcher hypothesized that addicted people would show a larger increase in brain activity following some treatment than would non-addicted people. In other words, the researcher hypothesized that addiction status and treatment would interact. The results shown in Figure $13.7.2$ are very much in line with the hypothesis. However, the test of the interaction resulted in a probability value of 0.08, a value not quite low enough to be significant at the conventional 0.05 level. The proper conclusion is that the experiment supports the researcher's hypothesis, but not strongly enough to allow a confident conclusion.

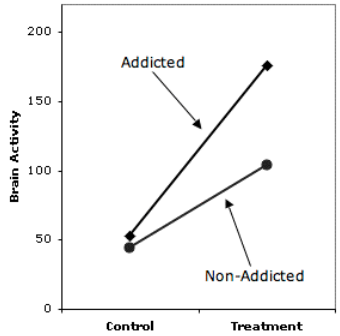


Figure 13.7.2: Made-up Data with One Significant Simple Effect

Unfortunately, the researcher was not satisfied with such a weak conclusion and went on to test the simple effects. It turned out that the effect of Treatment was significant for the Addicted group ($p = 0.02$) but not significant for the Non-Addicted group ($p = 0.09$). The researcher then went on to conclude that since there is an effect of Treatment for the Addicted group but not for the Non-Addicted group, the hypothesis of a greater effect for the former than for the latter group is demonstrated. This is faulty logic, however, since it is based on accepting the null hypothesis that the simple effect of Treatment is zero for the Non-Addicted group just because it is not significant.

Components of Interaction (optional)

Figure $13.7.3$ shows the results of an imaginary experiment on diet and weight loss. A control group and two diets were used for both overweight teens and overweight adults.

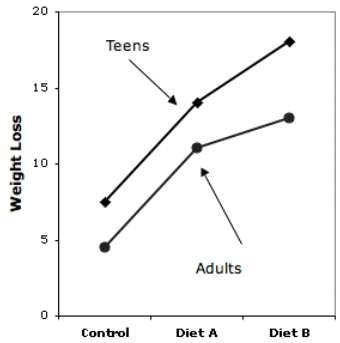


Figure 13.7.3: Made-up Data for Diet Study

The difference between Diet A and the Control diet was essentially the same for teens and adults, whereas the difference between Diet B and Diet A was much larger for the teens than it was for the adults. Over one portion of the graph the lines are parallel whereas over another portion they are not. It is possible to test these portions or components of interactions using the method of specific comparisons discussed previously. The test of the difference between Teens and Adults on the difference between Diets A and B could be tested with the coefficients shown in Table $13.7.4$ Naturally, the same considerations regarding multiple comparisons and orthogonal comparisons that apply to other comparisons among means also apply to comparisons involving components of interactions.

Table 13.7.4: Coefficients for a Component of the Interaction

Age Group	Diet	Coefficient
Teen	Control	0
Teen	A	1
Teen	B	-1
Adult	Control	0
Adult	A	-1
Adult	B	1

13.8: Within-Subjects

Learning Objectives

- Explain why a within-subjects design can be expected to have more power than a between-subjects design
- Explain error in terms of interaction
- Be able to create the Source and df columns of an ANOVA summary table for a design with one between-subjects and one within-subjects variable
- Describe the consequences of violating the assumption of sphericity
- Discuss courses of action that can be taken if sphericity is violated

Within-subjects factors involve comparisons of the same subjects under different conditions. For example, in the "ADHD Treatment" study, each child's performance was measured four times, once after being on each of four drug doses for a week.Therefore, each subject's performance was measured at each of the four levels of the factor "Dose." Note the difference from between-subjects factors for which each subject's performance is measured only once and the comparisons are among different groups of subjects. A within-subjects factor is sometimes referred to as a repeated-measures factor since repeated measurements are taken on each subject. An experimental design in which the independent variable is a within-subjects factor is called a within-subjects design.

Advantage of Within-Subjects Designs

One-Factor Designs

Let's consider how to analyze the data from the "ADHD Treatment" case study. These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. For now, we will be concerned only with testing the difference between the mean in the placebo condition (the lowest dosage, $D0$) and the mean in the highest dosage condition ($D60$). The details of the computations are relatively unimportant since they are almost universally done by computers. Therefore we jump right to the ANOVA Summary table shown in Table $13.8.1$.

Table 13.8.1: ANOVA Summary Table

Source	df	SSQ	MS	F	p
Subjects	23	5781.98	251.39		
Dosage	1	295.02	295.02	10.38	0.004
Error	23	653.48	28.41		
Total	47	6730.48			

The first source of variation, "Subjects," refers to the differences among subjects. If all the subjects had exactly the same mean (across the two dosages), then the sum of squares for subjects would be zero; the more subjects differ from each other, the larger the sum of squares subjects.

Dosage refers to the differences between the two dosage levels. If the means for the two dosage levels were equal, the sum of squares would be zero. The larger the difference between means, the larger the sum of squares.

The error reflects the degree to which the effect of dosage is different for different subjects. If subjects all responded very similarly to the drug, then the error would be very low. For example, if all subjects performed moderately better with the high dose than they did with the placebo, then the error would be low. On the other hand, if some subjects did better with the placebo while others did better with the high dose, then the error would be high. It should make intuitive sense that the less consistent the effect of dosage, the larger the dosage effect would have to be in order to be significant. The degree to which the effect of dosage differs depending on the subject is the $Subjects \times Dosage$ interaction. Recall that an interaction occurs when the effect of one variable differs depending on the level of another variable. In this case, the size of the error term is the extent to which the effect of the variable "Dosage" differs depending on the level of the variable "Subjects." Note that each subject is a different level of the variable "Subjects."

Other portions of the summary table have the same meaning as in between-subjects ANOVA. The F for dosage is the mean square for dosage divided by the mean square error. For these data, the F is significant with $p = 0.004$. Notice that this F test is equivalent to the t test for correlated pairs, with $F = t^2$.

Table $13.8.2$ shows the ANOVA Summary Table when all four doses are included in the analysis. Since there are now four dosage levels rather than two, the df for dosage is three rather than one. Since the error is the $Subjects \times Dosage$ interaction, the df for error is the df for "Subjects" (23) times the df for Dosage (3) and is equal to 69.

Table 13.8.2: ANOVA Summary Table

Source	df	SSQ	MS	F	p
Subjects	23	9065.49	394.15		
Dosage	3	557.61	185.87	5.18	0.003
Error	69	2476.64	35.89		
Total	95	12099.74			

Carryover Effects

Often performing in one condition affects performance in a subsequent condition in such a way as to make a within-subjects design impractical. For example, consider an experiment with two conditions. In both conditions subjects are presented with pairs of words. In Condition A, subjects are asked to judge whether the words have similar meaning whereas in Condition B, subjects are asked to judge whether they sound similar. In both conditions, subjects are given a surprise memory test at the end of the presentation. If Condition were a within-subjects variable, then there would be no surprise after the second presentation and it is likely that the subjects would have been trying to memorize the words.

Not all carryover effects cause such serious problems. For example, if subjects get fatigued by performing a task, then they would be expected to do worse on the second condition they were in. However, as long as the order of presentation is counterbalanced so that half of the subjects are in Condition A first and Condition Bsecond, the fatigue effect itself would not invalidate the results, although it would add noise and reduce power. The carryover effect is symmetric in that having Condition A first affects performance in Condition B to the same degree that having Condition B first affects performance in Condition A.

Asymmetric carryover effects cause more serious problems. For example, suppose performance in Condition B were much better if preceded by Condition A, whereas performance in Condition A was approximately the same regardless of whether it was preceded by Condition B. With this kind of carryover effect, it is probably better to use a between-subjects design.

One between- and one within-subjects factor

In the "Stroop Interference" case study, subjects performed three tasks: naming colors, reading color words, and naming the ink color of color words. Some of the subjects were males and some were females. Therefore, this design had two factors: gender and task. The ANOVA Summary Table for this design is shown in Table $13.8.3$

Table 13.8.3: ANOVA Summary Table for Stroop Experiment

Source	df	SSQ	MS	F	p
Gender	1	83.32	83.32	1.99	0.165
Error	45	1880.56	41.79		
Task	2	9525.97	4762.99	228.06	<0.001
Gender x Task	2	55.85	27.92	1.34	0.268
Error	90	1879.67	20.89		

The computations for the sums of squares will not be covered since computations are normally done by software. However, there are some important things to learn from the summary table. First, notice that there are two error terms: one for the between-subjects variable Gender and one for both the within-subjects variable Task and the interaction of the between-subjects variable and the

within-subjects variable. Typically, the mean square error for the between-subjects variable will be higher than the other mean square error. In this example, the mean square error for Gender is about twice as large as the other mean square error.

The degrees of freedom for the between-subjects variable is equal to the number of levels of the between-subjects variable minus one. In this example, it is one since there are two levels of gender. Similarly, the degrees of freedom for the within-subjects variable is equal to the number of levels of the variable minus one. In this example, it is two since there are three tasks. The degrees of freedom for the interaction is the product of the degrees of freedom for the two variables. For the $Gender \times Task$ interaction, the degrees of freedom is the product of degrees of freedom Gender (which is 1) and the degrees of freedom Task (which is 2) and is equal to 2.

Assumption of Sphericity

Within-subjects ANOVA makes a restrictive assumption about the variances and the correlations among the dependent variables. Although the details of the assumption are beyond the scope of this book, it is approximately correct to say that it is assumed that all the correlations are equal and all the variances are equal. Table $13.8.4$ shows the correlations among the three dependent variables in the "Stroop Interference" case study.

Table $13.8.4$: Correlations Among Dependent Variables

	word reading	color naming	interference
word reading	1	0.7013	0.1583
color naming	0.7013	1	0.2382
interference	0.1583	0.2382	1

Note that the correlation between the word reading and the color naming variables of 0.7013 is much higher than the correlation between either of these variables with the interference variable. Moreover, as shown in Table $13.8.5$, the variances among the variables differ greatly.

Table $13.8.5$: Variances

Variable	Variance
word reading	15.77
color naming	13.92
interference	55.07

Naturally the assumption of sphericity, like all assumptions, refers to populations not samples. However, it is clear from these sample data that the assumption is not met in the population.

Consequences of Violating the Assumption of Sphericity

Although ANOVA is robust to most violations of its assumptions, the assumption of sphericity is an exception: Violating the assumption of sphericity leads to a substantial increase in the Type I error rate. Moreover, this assumption is rarely met in practice. Although violations of this assumption had at one time received little attention, the current consensus of data analysts is that it is no longer considered acceptable to ignore them.

Approaches to Dealing with Violations of Sphericity

If an effect is highly significant, there is a conservative test that can be used to protect against an inflated **Type I** error rate. This test consists of adjusting the degrees of freedom for all within-subjects variables as follows: The degrees of freedom numerator and denominator are divided by the number of scores per subject minus one. Consider the effect of Task shown in Table $13.8.3$. There are three scores per subject and therefore the degrees of freedom should be divided by two. The adjusted degrees of freedom are:

$(2)(1/2) = 1$ for the numerator and

$(90)(1/2) = 45$ for the denominator

The probability value is obtained using the F probability calculator with the new degrees of freedom parameters. The probability of an F of 228.06 or larger with 1 and 45 degrees of freedom is less than 0.001. Therefore, there is no need to worry about the assumption violation in this case.

Possible violation of sphericity does make a difference in the interpretation of the analysis shown in Table $13.8.2$. The probability value of an F of 5.18 with 1 and 23 degrees of freedom is 0.032, a value that would lead to a more cautious conclusion than the p value of 0.003 shown in Table $13.8.2$.

The correction described above is very conservative and should only be used when, as in Table $13.8.3$, the probability value is very low. A better correction, but one that is very complicated to calculate, is to multiply the degrees of freedom by a quantity called ε (the Greek letter epsilon). There are two methods of calculating ε. The correction called the Huynh-Feldt (or $H-F$) is slightly preferred to the one called the Greenhouse-Geisser (or $G-G$), although both work well. The $G-G$ correction is generally considered a little too conservative.

A final method for dealing with violations of sphericity is to use a multivariate approach to within-subjects variables. This method has much to recommend it, but it is beyond the scope of this text.

13.9: Power of Within-Subjects Designs Demo

Learning Objectives

- State the relationship between variance and power
- State the effect of using a one-tailed test on power

Instructions

This simulation demonstrates the effect of the correlation between measures in a one-way within-subjects ANOVA with two levels. This test is equivalent to a correlated t test. The default values for this demonstration are for an experiment with 10 subjects each measured under two conditions. The population difference for the two conditions is 1.85 and the variance in each of the conditions is 4.0. The graph shows the power of the test as a function of the population correlation between the two scores for the 0.10, 0.05, and 0.01 significance levels. The power of an independent-groups t test (which assumes the correlation is 0) is shown by the $x's$.

Experiment with different combinations of the parameters. Is the correlation an important factor in power?

Illustrated Instructions

Video Demo

The video begins by changing the population variance to 8 and increases the sample size to 25 and then reduces it to 5. Notice the impact that these changes have on the relationship between power and population correlation. The video concludes by changing the mean difference to 4.

13.10: Statistical Literacy

Learning Objectives

- Testing an Alzheimer's Drug

A research design to compare three drugs for the treatment of Alzheimer's disease is described here. For the first two years of the study, researchers will follow the subjects with scans and memory tests.

Example 13.10.1: what do you think?

Assume the data were analyzed as a two-factor design with pre-post testing as one factor and the three drugs as the second factor. What term in an ANOVA would reflect whether the pre-post change was different for the three drugs?

Solution

It would be the interaction of the two factors since the question is whether the effect of one factor (pre-post) differs as a function of the level of a second factor (drug).

Weight Loss and Cancer Risk

An experiment on risk factors for cancer had four conditions related to weight loss. Quoting the web page description:

> "McTiernan and her colleagues randomly assigned 399 overweight and obese postmenopausal women to one of four groups — dieting plus exercise, dieting only, exercise only or a control group not required to do either."

Example 13.10.2: what do you think?

What type of ANOVA design was used?

Solution

There are two between-subjects factors: Diet (yes or no) and Exercise (yes or no) resulting in the four conditions. The design is a 2×2 factorial design.

13.E: Analysis of Variance (Exercises)

General Questions

Q1

What is the null hypothesis tested by analysis of variance?

Q2

What are the assumptions of between-subjects analysis of variance?

Q3

What is a between-subjects variable?

Q4

Why not just compute t-tests among all pairs of means instead computing an analysis of variance?

Q5

What is the difference between "N" and "n"?

Q6

How is it that estimates of variance can be used to test a hypothesis about means?

Q7

Explain why the variance of the sample means has to be multiplied by "n" in the computation of MSB.

Q8

What kind of skew does the F distribution have?

Q9

When do MSB and MSE estimate the same quantity?

Q10

If an experiment is conducted with 6 conditions and 5 subjects in each condition, what are dfn and dfe?

Q11

How is the shape of the F distribution affected by the degrees of freedom?

Q12

What are the two components of the total sum of squares in a one-factor between-subjects design?

Q13

How is the mean square computed from the sum of squares?

Q14

An experimenter is interested in the effects of two independent variables on self esteem. What is better about conducting a factorial experiment than conducting two separate experiements, one for each independent variable?

Q15

An experiment is conducted on the effect of age and treatment condition (experimental versus control) on reading speed. Which statistical term (main effect, simple effect, interaction, specific comparison) applies to each of the descriptions of effects.

a. The effect of the treatment was larger for 15-year olds than it was for 5- or 10-year olds.
b. Overall, subjects in the treatment condition performed faster than subjects in the control condition.
c. The difference between the 10- and 15-year olds was significant under the treatment condition.

d. The difference between the 15- year olds and the average of the 5- and 10-year olds was significant.
e. As they grow older, children read faster.

Q16

An $A(3) \times B(4)$ factorial design with 6 subjects in each group is analyzed. Give the source and degrees of freedom columns of the analysis of variance summary table.

Q17

The following data are from a hypothetical study on the effects of age and time on scores on a test of reading comprehension. Compute the analysis of variance summary table.

	12-year olds	16-year olds
30 minutes	66 68 59 72 46	74 71 67 82 76
60 minutes	69 61 69 73 61	95 92 95 98 94

Q18

Define "Three-way interaction"

Q19

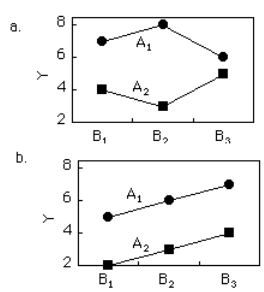
Define interaction in terms of simple effects.

Q20

Plot an interaction for an $A(2) \times B(2)$ design in which the effect of B is greater at $A1$ than it is at $A2$. The dependent variable is "Number correct." Make sure to label both axes.

Q21

Following are two graphs of population means for 2×3 designs. For each graph, indicate which effect(s) (A, B, or $A \times B$) are nonzero.



Q22

The following data are from an $A(2) \times B(4)$ factorial design.

	B1	B2	B3	B4
A1	1 3 4 5	2 2 4 5	3 4 2 6	4 5 6 8
A2	1 1 2 2	2 3 2 4	4 6 7 8	8 9 9 8

a. Compute an analysis of variance.
b. Test differences among the four levels of B using the Bonferroni correction.
c. Test the linear component of trend for the effect of B.
d. Plot the interaction.
e. Describe the interaction in words.

Q23

Why are within-subjects designs usually more powerful than between-subjects design?

Q24

What source of variation is found in an ANOVA summary table for a within-subjects design that is not in in an ANOVA summary table for a between-subjects design. What happens to this source of variation in a between-subjects design?

Q25

The following data contain three scores from each of five subjects. The three scores per subject are their scores on three trials of a memory task.

$$\begin{array}{ccc} 4 & 6 & 7 \\ 3 & 7 & 7 \\ 2 & 8 & 5 \\ 1 & 4 & 7 \\ 4 & 6 & 9 \end{array} \tag{13.E.1}$$

a. Compute an ANOVA
b. Test all pairwise differences between means using the Bonferroni test at the 0.01 level.
c. Test the linear and quadratic components of trend for these data.

Q26

Give the source and df columns of the ANOVA summary table for the following experiments:

a. 22 subjects are each tested on a simple reaction time task and on a choice reaction time task.
b. 12 male and 12 female subjects are each tested under three levels of drug dosage: $0mg, 10mg, 20mg$
c. 20 subjects are tested on a motor learning task for 3 trials a day for 2 days.
d. An experiment is conducted in which depressed people are either assigned to a drug therapy group, a behavioral therapy group, or a control group. 10 subjects are assigned to each group. The level of measured once a month for 4 months.

Questions from Case Studies

The following question is from the Stroop Interference case study.

Q27

The dataset has the scores (times) for males and females on each of three tasks.

a. Do a $Gender(2) \times Task(3)$ analysis of variance.
b. Plot the interaction.

The following question is from the ADHD Treatment case study.

Q28

The data has four scores per subject.

a. Is the design between-subjects or within-subjects?
b. Create an ANOVA summary table.

The following question is from the Angry Moods case study.

Q29

Using the Anger Expression Index as the dependent variable, perform a 2×2 ANOVA with gender and sports participation as the two factors. Do athletes and non-athletes differ significantly in how much anger they express? Do the genders differ significantly in Anger Expression Index? Is the effect of sports participation significantly different for the two genders?

The following question is from the Weapons and Aggression case study.

Q30

Compute a 2×2 ANOVA on this data with the following two factors: prime type (was the first word a weapon or not?) and word type (was the second word aggressive or non-aggressive?). Consider carefully whether the variables are between-subject or within-subects variables.

The following question is from the Smiles and Leniency case study.

Q31

Compute the ANOVA summary table.

CHAPTER OVERVIEW

14: Case Studies

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

14.1: Angry Moods

Learning Objectives

- Ways to improve an angry mood: A look at gender and sports participation

Research conducted by

Emily Zitek and Mindy Ater, Rice University

Case study prepared by

Emily Zitek

Overview

People have different ways of improving their mood when angry. We have all seen people punch a wall when mad, and indeed, previous research has indicated that some people aggress to improve their mood (Bushman, Baumeister & Phillips, 2001). What do the top athletes do when angry? Striegel (1994) found that anger often hurts an athlete's performance and that capability to control anger is what makes good athletes even better. This study adds to the past research and examines the difference in ways to improve an angry mood by gender and sports participation.

The participants were 78 Rice University undergraduates, ages 17 to 23. Of these 78 participants, 48 were females and 30 were males and 25 were athletes and 53 were non-athletes. People who did not play a varsity or club sport were considered non-athletes. The 13 contact sport athletes played soccer, football, rugby, or basketball, and the 12 non-contact sport athletes participated in Ultimate Frisbee, baseball, tennis, swimming, volleyball, crew, or dance.

The participants were asked to respond to a questionnaire that asked about what they do to improve their mood when angry or furious. Then they filled out a demographics questionnaire.

Note

This study used the most recent version of the State-Trait Anger Expression Inventory (STAXI-2) (Spielberger, Sydeman, Owen & Marsh, 1999) which was modified to create an Angry Mood Improvement Inventory similar to that created by Bushman et al. (2001).

Questions to Answer

Do athletes and non-athletes deal with anger in the same way? Are there any gender differences? Specifically, are men more likely to believe that aggressive behavior can improve an angry mood?

Design Issues

This study has an extremely unbalanced design. There were a lot more non-athletes than athletes in the sample. In the future, more athletes should be used. This study originally wanted to look at contact and non-contact athletes separately, but there were not enough participants to do this. Future studies could look at this.

Descriptions of Variables

Table 14.1.1: Description of Variables. Note that the description of the items comes from Spielberger et al. (1999)

Variable	Description
Sports	1 = athletes, 2 = non-athletes
Gender	1 = males, 2 = females

Anger-Out (AO)	high scores demonstrate that people deal with anger by expressing it in a verbally or physically aggressive fashion
Anger-In (AI)	high scores demonstrate that people experience anger but do not express it (suppress their anger)
Control-Out (CO)	high scores demonstrate that people control the outward expression of angry feelings
Control-In (CI)	high scores demonstrate that people control angry feelings by calming down or cooling off
Expression (AE)	index of general anger expression: (Anger-Out) + (Anger-In) - (Control-Out) - (Control-In) + 48

Data files

angry_moods.xls

References

- Bushman, B.J., Baumeister, R.F. & Phillips, C.M. (2001). Do people aggress to improve their mood? Catharsis beliefs, affect regulation opportunity, and aggressive responding. Journal of Personality and Social Psychology, 81(1), 17-32.
- Spielberger, C. D., Sydeman, S. J., Owen, A. E., Marsh, B. J. (1999). Measuring anxiety and anger with the State-Trait Anxiety Inventory (STAI) and the State-Trait Anger Expression Inventory (STAXI). In M. E. Maruish (Ed.), The use of psychological testing for treatment planning and outcomes assessment (2nd ed., pp. 993-1021). Mahwah: Lawrence Erlbaum Associates.
- Striegel, D. (1994). Anger in tennis: Part 2. Effects of anger on performance, coping with anger, and using anger to one's benefit. Journal of Performance Psychology, 2, 56-92.

Links

Publisher's description

14.2: Flatulence

Learning Objectives

- Flatulence: Are you embarrassed by your flatus?

Research conducted by

Shannon E. Collins, UH-D undergraduate,

Faculty Advisor: Heidi Ziemer

Case study prepared by

Shannon E. Collins

Overview

The purpose of this study was to find out whether or not people are embarrassed by their flatulence. The participants were 35 University of Houston – Downtown students. Flatulence is a normal part of being human, but it can cause an alarming rate of embarrassment in certain situations. How many times have you been subjected to the unpleasant odor emitted from someone around you? Medical research indicates that it is normal to have anywhere from 7 to 20 episodes of gas in a day.

Would you believe that women produce more of the bad smelling stuff than men do, and that women are more likely to complain to their doctors about the smell of their flatulence? The smell comes from sulfur gasses, the most offensive of which is hydrogen sulfide; it smells like rotten eggs. Still, everybody does it, we just don't know how embarrassed they are by it.

Questions to Answer

Are people without male siblings more embarrassed by their flatulence than people with one or more male siblings? Do people that come from households where flatulence was acceptable report less embarrassment than people that come from households where it was not acceptable? Are women or men more embarrassed by their flatus?

Design Issues

Embarrassment scores were reported on 14 different measures and tallied as a total embarrassment score, then divided into seven categories, producing one number per category. Only the scores on the seven categories are reported here. The data in this research is self report data, and because the topic is sensitive some people may have been less than honest about their reported flatulence.

Descriptions of Variables

Table 14.2.1: Description of variables

Variable	Description
Gender	1 = male, 2 = female
famaccp	Household acceptance of flatus 1 to 7, 1=very acceptable and 7=very unacceptable
brother	Number of brothers the participant has
howlong	How long before farting in front of partner? 1 = 1 year, .5= 6 months, .25=3 months, and smaller decimals represent portions of a year. Numbers larger than 1 indicate longer than 1 year.
perday	Number per day

Embarrassing Situations	The following variables were rated on this scale: 1=extremely embarrassed and 7= not really embarrassed
mtgwork	Meeting at work
talkprof	Talking to a professor
romint	Romantic interest

Data files

flatulence.xls

References

- Chapman, S. (2001 December 22). Hot Air? BMJ: British Medical Journal. 323(7327). Retrieved from, Health source database.
- Gases of the Gut. Harvard Health Letter. August 2002. 27(10).
- Hughes, L. (October 1999). 10 Tips on How to Fend Off Embarrassing Flatulence. Environmental Nutrition. 22(10). Retrieved from, Health source database.
- Lecture Theatre Etiquette. (June 2000). Student BMJ. Vol. 8. Retrieved from, Health source database.
- Manley, W. (October 1999). Noisome Rumblings. American Libraries Online. Retrieved from, www.search.epnet.com/direct.a...4&db=hch&tg=AN
- Robb-Nicholson, C. (March 2003). By The Way Doctor. Harvard Women's Health Watch. 10(7). Retrieved from, Academic Edition of Health Source database.

14.3: Physicians Reactions

Learning Objectives

- Physicians' Reactions to Patient Size

Research conducted by

Mikki Hebl and Jingping Xu

Case study prepared by

Emily Zitek

Overview

Obese people face discrimination on a daily basis in employment, education, and relationship contexts. Past research has shown that even doctors, who are trained to treat all their patients warmly and have access to literature suggesting uncontrollable and hereditary aspects of obesity, believe obese individuals are undisciplined and suffer from controllability issues. This case study examines how doctors treat overweight as compared to normal weight patients.

Various doctors at one of three major hospitals in the Texas Medical Center of Houston participated in the study. These doctors were sent a packet containing a medical chart similar to the one they view upon seeing a patient. This chart portrayed a patient who was displaying symptoms of a migraine headache but was otherwise healthy. This chart also contained a measure of the patient's weight. Doctors were randomly assigned to receive the chart of a patient who was overweight or the chart of a patient who was of normal weight. After reviewing the chart, the doctors then had to indicate how much time they believed they would spend with the patient.

Questions to Answer

Do doctors discriminate against overweight patients? Specifically, do the doctors who review charts of overweight patients say they would spend the same amount of time with their patients as the doctors who review charts of normal weight patients?

Design Issues

The method and data described here are only a small part of a larger study. See the reference below for a full description of the study.

Descriptions of Variables

Table 14.3.1: Description of variables

Variable	Description
Patient weight	1 = average weight, 2 = overweight
Time	represents how long the doctors said they would spend with the patient

Data files

Weight.xls

References

- Hebl, M., & Xu, J., "Weighing the care: Physicians' reactions to the size of a patient," International Journal of Obesity, 25 (2001): 1246-1252

14.4: Teacher Ratings

Learning Objectives

- Teacher Ratings

Research conducted by

Annette Towler and Robert Dipboye

Case study prepared by

Emily Zitek

Overview

How powerful are rumors? Frequently, students ask friends and/or look at instructor evaluations to decide if a class is worth taking. Kelley (1950) found that instructor reputation has a profound impact on actual teaching ratings, and Towler and Dipboye (1998) replicated and extended this study.

Subjects were randomly assigned to one of two conditions. Before viewing the lecture, students were given a summary of the instructors' prior teaching evaluations. There were two conditions: Charismatic instructor and Punitive instructor.

Then all subjects watched the **same** twenty-minute lecture given by the exact same lecturer. Following the lecture, subjects answered three questions about the leadership qualities of the lecturer. A summary rating score was computed and used as the variable "rating" here.

Questions to Answer

Does an instructor's prior reputation affect student ratings?

Design Issues

The data presented here are part of a larger study. See the references below to learn more.

Descriptions of Variables

Table 14.4.1: Description of Variables

Variable	Description
Condition	this represents the content of the description that the students were given about the professor (1 = charismatic, 2 = punitive)
Rating	how favorably the subjects rated the professor after hearing the lecture (higher ratings are more favorable)

Data files

Ratings.xls

References

- Kelley, H. H.(1950). The warm-cold variable in first impression of persons. Journal of Personality, 18, 431-439.
- Towler, A., & Dipboye, R. L. (1998). The effect of instructor reputation and need for cognition on student behavior (poster presented at American Psychological Society conference, May 1998).

14.5: Diet and Health

Learning Objectives

- Mediterranean Diet and Health

Research conducted by

De Longerill et al

Case study prepared by

David Lane and Emily Zite

Overview

Most doctors would probably agree that a Mediterranean diet, rich in vegetables, fruits, and grains, is healthier than a high-saturated fat diet. Indeed, previous research has found that the diet can lower risk of heart disease. However, there is still considerable uncertainty about whether the Mediterranean diet is superior to a low-fat diet recommended by the American Heart Association. This study is the first to compare these two diets.

The subjects, 605 survivors of a heart attack, were randomly assigned follow either

1. a diet close to the "prudent diet step 1" of the American Heart Association (control group) or
2. a Mediterranean-type diet consisting of more bread and cereals, more fresh fruit and vegetables, more grains, more fish, fewer delicatessen foods, less meat.

An experimental canola-oil-based margarine was used instead of butter or cream. The oils recommended for salad and food preparation were canola and olive oils exclusively. Moderate red wine consumption was allowed.

Over a four-year period, patients in the experimental condition were initially seen by the dietician, two months later, and then once a year. Compliance with the dietary intervention was checked by a dietary survey and analysis of plasma fatty acids. Patients in the control group were expected to follow the dietary advice given by their physician.

The researchers collected information on number of deaths from cardiovascular causes e.g., heart attack, strokes, as well as number of nonfatal heart-related episodes. The occurrence of malignant and nonmalignant tumors was also carefully monitored.

Questions to Answer

Is the Mediterranean diet superior to a low-fat diet recommended by the American Heart Association?

Design Issues

The strength of the design is that subjects were randomly assigned to conditions. A possible weakness is that compliance rates depended on reports rather than observation since observation is impractical in this type of research.

Descriptions of Variables

Table $14.5.1$: Description of Variables

Variable	Description
Type of diet	AHA or Mediterranean
Various outcome measures of health and disease	does the patient have cancer, etc.?

Data Files

Diet.xls

Links

More on the Mediterranean Diet

References

- De Longerill, M., Salen, P., Martin, J., Monjaud, I., Boucher, P., Mamelle, N. (1998). Mediterranean Dietary pattern in a Randomized Trial. Archives of Internal Medicine, 158, 1181-1187.

14.6: Smiles and Leniency

Learning Objectives

- To study the research on effects of smiling

Research conducted by

Marianne LaFrance and Marvin Hecht

Case study prepared by

David Lane

Overview

Dale Carnegie stated that smiling helps win friends and influence people. Research on the effects of smiling has backed this up and shown that a smiling person is judged to be more pleasant, attractive, sincere, sociable, and competent than a non-smiling person.

There is evidence that smiling can attenuate judgments of possible wrongdoing. This phenomenon termed the "smile-leniency effect" was the focus of a study by Marianne LaFrance & Marvin Hecht in 1995.

Questions to Answer

Does smiling increase leniency? Are different types of smiles differentially effective?

Design Issues

There was a single person used for all the conditions. This may limit the generalizeability of the results.

Descriptions of Variables

Table 14.6.1 : Description of Variables

Variable	Description	
Smile	1 is false smile 2 is felt smile 3 is miserable smile 4 is neutral control	
Leniency	A measure of how lenient the judgments were.	

Data Files

Leniency.xls

References

- LaFrance, M., & Hecht, M. A. (1995) Why smiles generate leniency. Personality and Social Psychology Bulletin, 21, 207-214.

14.7: Animal Research

Learning Objectives

- Gender difference in attitudes toward the use of animals in research

Research conducted by

Nicole Hilliard, Faculty Advisor: Heidi Ziemer

Case study prepared by

Emily Zitek

Overview

The use of animals in research is a controversial and emotionally charged issue. Personal feelings regarding the use of animals in research vary widely. While many believe that the use of animals in research has been and continues to be essential, others want the practice stopped by cutting off funding or the passing of legislative restrictions. Research on human attitudes toward the use of animals in research has consistently shown systematic differences of opinion with gender differences among the largest.

In this study, a convenience sample of 34 University of Houston - Downtown students completed a simple survey that asked their gender and how much they agreed with the following two statements: "The use of animals in research is wrong," and "The use of animals in research is necessary". They rated their agreement with each of these statements on a 7-point scale from strongly disagree (1) to strongly agree (7).

Questions to Answer

Is there a gender difference with respect to the belief that animal research is wrong? Is there a gender difference with respect to the belief that animal research is necessary?

Design Issues

This is self-report data. It is possible that the willingness to admit to thinking animal research is wrong or necessary is what differs by gender, not how the participants actually feel.

Descriptions of Variables

Table 14.7.1: Description of Variables

Variable	Description
Gender	1 = female, 2 = male
Wrong	high scores indicate that the participant believes that animal research is wrong
Necessary	high scores indicate that the participant believes that animal research is necessary

Data Files

Animals.xls

Links

American Association for the Advancement of Science

References

- Eldridge, J.J. & Gluck, J.P. (1996) Gender differences in attitudes toward animal research. Ethics & Behavior, 6(3), 239-256.
- Nickell, D & Herzog, H.A. (1996). Ethical ideology and moral persuasion: Personal moral philosophy, gender, and judgements of pro- and anti-animal research propaganda. Society & Animals, 4(1), 53-64.
- Pifer, L. K. (1996). Exploring the gender gap in young adults' attitudes about animal research. Society & Animals, 4(1), 37-52.
- Wuensch, K. L. & Poteat, G.M. (1998). Evaluating the morality of animal research: Effects of ethical ideology, gender, and purpose. *Journal of Social Behavior & Personality*, *13*(1), 139-151.

14.8: ADHD Treatment

Learning Objectives

- Treatment Effects of a Drug on Cognitive Functioning in Children with Mental Retardation and ADHD

Research conducted by

Pearson et al. (2003, see reference below)

Case study prepared by

David Lane and Emily Zitek

Overview

This study investigated the cognitive effects of stimulant medication in children with mental retardation and Attention-Deficit/Hyperactivity Disorder. This case study shows the data for the Delay of Gratification (DOG) task. Children were given various dosages of a drug, methylphenidate (MPH) and then completed this task as part of a larger battery of tests. The order of doses was counterbalanced so that each dose appeared equally often in each position. For example, six children received the lowest dose first, six received it second, etc. The children were on each dose one week before testing.

This task, adapted from the preschool delay task of the Gordon Diagnostic System (Gordon, 1983), measures the ability to suppress or delay impulsive behavioral responses. Children were told that a star would appear on the computer screen if they waited "long enough" to press a response key. If a child responded sooner in less than four seconds after their previous response, they did not earn a star, and the 4-second counter restarted. The DOG differentiates children with and without ADHD of normal intelligence (e.g., Mayes et al., 2001), and is sensitive to MPH treatment in these children (Hall & Kataria, 1992).

Questions to Answer

Does higher dosage lead to higher cognitive performance (measured by the number of correct responses to the DOG task)?

Design Issues

This is a repeated-measures design because each participant performed the task after each dosage.

Descriptions of Variables

Table 14.8.1: Description of Variables

Variable	Description
d0	Number of correct responses after taking a placebo
d15	Number of correct responses after taking .15 mg/kg of the drug
d30	Number of correct responses after taking .30 mg/kg of the drug
d60	Number of correct responses after taking .60 mg/kg of the drug

Data Files

ADHD.xls

Links

Methylphenidate

References

- Gordon M (1983), The Gordon Diagnostic System. DeWitt, NY: Gordon Systems
- Hall CW, Kataria S (1992), Effects of two treatment techniques on delay and vigilance tasks with attention deficit hyperactive disorder (ADHD) children. J Psychol 126:17-25
- Mayes SD, Calhoun SL, Crowell, EW (2002), The Gordon Diagnostic System and WISC-III Freedom from Distractibility index: Validity in identifying clinic-referred children with and without ADHD. Psychol Rep ,91, 575-587.
- Pearson DA, Santos CW, Roache JD, Casat CD, Loveland KA, Lachar D, Lane DM, Faria, LF, Cleveland LA (2003), Treatment effects of methylphenidate on behavioral adjustment in children with mental retardation and ADHD. J Am Acad Child Adolesc Psychiatry, 42, 209-216.
- Pearson, D.A., Santos, C.W., Jerger, S.W., Casat, C.D., Roache, J., Loveland, K.A., Lane, D.M., Lachar, D., Faria, L.P., & Getchell, C. (2003) Treatment effects of methylphenidate on cognitive functioning in children with mental retardation and ADHD. Journal of the American Academy of Child and Adolescent Psychiatry, 43, 677-685.

14.9: Weapons and Aggression

Learning Objectives

- Study of the "Weapons" effect

Research conducted by

Anderson, Benjamin, and Bartholow

Case study prepared by

David Lane

Overview

The "weapons effect" is the finding that the presence of a weapon or even a picture of a weapon can cause people to behave more aggressively. Although once a controversial finding, the weapons effect is now a well-established phenomenon. Based on this, Anderson, Benjamin, and Bartholow (1998) hypothesize that the presence of a weapon-word prime (such as "dagger" or "bullet") should increase the accessibility of an aggressive word (such as "destroy" or "wound"). The accessibility of a word can be measured by the time it takes to name a word presented on computer screen.

The subjects were undergraduate students ranging in age from 18 to 24 years. They were told that the purpose of this study was to test reading ability of various words. On each of the 192 trials, a computer presented a priming stimulus word (either a weapon or non-weapon word) for 1.25 seconds, a blank screen for 0.5 seconds, and then a target word (aggressive or non-aggressive word). Each subject named both aggressive and non-aggressive words following both weapon and non-weapon "primes." The experimenter instructed the subjects to read the first word to themselves and then to read the second word out loud as quickly as they could. The computer recorded response times and computed mean response times for each participant for each of the four conditions.

Examples of the four types of words

- Weapon word primes: shotgun, grenade
- Non-weapon word primes: rabbit, fish
- Aggressive word: injure, shatter
- Non-aggressive word: consider, relocate

Questions to Answer

Does the mere presence of a weapon increase the accessibility of aggressive thoughts? More specifically, can a person name an aggressive word more quickly if it is preceded by a weapon word prime than if it is preceded by a neutral (non-aggressive) word prime?

Design Issues

This is a within-subjects design, and each participant provided four scores to the analysis.

Descriptions of Variables

Table $14.9.1$: Description of Variables

Variable	Description
gender	1 = female, 2 = male
aw	The time in milliseconds (msec) to name aggressive word following a weapon word prime.

an	The time in milliseconds (msec) to name aggressive word following a non-weapon word prime.
cw	The time in milliseconds (msec) to name a control word following a weapon word prime.
cn	The time in milliseconds (msec) to name a control word following a non-weapon word prime.

Data Files

Guns.xls

References

- Anderson, C.A., Benjamin, A.J., & Bartholow, B.D. (1998). Does the gun pull the trigger? Automatic priming effects of weapon pictures and weapon names. Psychological Science, 9, 308-314.

14.10: SAT and College GPA

Learning Objectives

- Predicting college GPA from high school scores

Research conducted by

Thomas W. MacFarland

Case study prepared by

Emily Zitek

Overview

When deciding whether to admit an applicant, colleges take lots of factors, such as grades, sports, activities, leadership positions, awards, teacher recommendations, and test scores, into consideration. Using SAT scores as a basis of whether to admit a student or not has created some controversy. Among other things, people question whether the SATs are fair and whether they predict college performance.

This study examines the SAT and GPA information of 105 students who graduated from a state university with a B.S. in computer science. Using the grades and test scores from high school, can you predict a student's college grades?

Questions to Answer

Can the math and verbal SAT scores be used to predict college GPA? Are the high school and college GPAs related?

Design Issues

The conclusions from this study should not be generalized to students of other majors.

Descriptions of Variables

Table 14.10.1: Description of Variables

Variable	Description
high_GPA	High school grade point average
math_SAT	Math SAT score
verb_SAT	Verbal SAT score
comp_GPA	Computer science grade point average
univ_GPA	Overall university grade point average

Data Files

SAT.xls

Links

Want a job? Hand over your SAT results! Is the SAT a fair test?

References

- None

14.11: Stereograms

Learning Objectives

- Study to determine the effects of information for an embedded image given ahead of time to a person

Research conducted by

Frisby, J. P. and Clatworthy, J.L.

Case study prepared by

Emily Zitek from DASL story contributed by Michael Friendly

Overview

The rectangles below appear to be composed of random dots. However, if the images are viewed with a stereo viewer, the separate images will fuse and reveal an embedded $3D$ figure. In this example, fusing the images of these random dot stereograms will reveal a diamond. (Another way for you to fuse the images is to fixate on a point in between them and defocus your eyes. This technique takes practice, but you can try it out with the links below.)

This experiment sought to determine whether giving someone information about the embedded image can help speed up how long it takes to view this image. Seventy-eight participants were given no information, verbal information, and/or visual information (a drawing of the object) about what the embedded image should look like before attempting to fuse the images and actually view the 3D design.

Figure 14.11.1: Random dots form an embedded image when viewed with a stereo viewer

Questions to Answer

Does giving someone information about an embedded image in a stereogram affect the amount of time it takes to see this image? More specifically, does the amount of time it takes to fuse the image in a stereogram differ when the person is given both verbal and visual information about what the image should look like as opposed to when the person is only given verbal information or no information at all?

Descriptions of Variables

Table 14.11.1: Description of Variables

Variable	Description
Time	Time to produce a fused image of the random dot stereogram
Group	Treatment group divided by type of information received: 1 = no information or only verbal information 2 = both verbal and visual information

Data Files

Fusion.xls

Links

View random dot stereograms. Information about random dot stereograms

References

- Frisby, J. P. & Clatworthy, J.L., (1975) Learning to see complex random-dot stereograms, Perception, 4, 173-178.

14.12: Driving

Learning Objectives

- Driving in inclement weather



Research conducted by

Darin Baskin

Case study prepared by

Emily Zitek

Overview

Many people believe that weather patterns influence driving safety. As a result, there are many web sites and other publications dedicated to giving people tips about how to drive in various weather conditions (see references and links below). Additionally, car accidents are often attributed to bad weather (e.g., see Taylor & Quinn, 1991). This study examines the beliefs and behaviors of people with respect to the important topic of driving in inclement weather.

The participants in this study filled out a questionnaire consisting of some demographic questions and then questions asking about their transportation habits and other beliefs concerning inclement weather. This questionnaire was administered to a convenience sample of 61 University of Houston - Downtown students at various locations (i.e., classrooms, hallways, and the food court).

Questions to Answer

Is gender or age related to the likelihood of driving in inclement weather? Does the number of accidents that someone thinks occur during inclement weather relate to how often he or she takes public transportation or chooses to drive during inclement weather?

Design Issues

This is a correlational study, so we cannot infer causation.

Descriptions of Variables

Variable	Description
Age	The age of the participant in years
Gender	1 = female, 2 = male
Cho2drive	How often he or she chooses to drive in inclement weather 1 = always, 3 = sometimes, 5 = never

Pubtran	% of travel time spent on public transportation in inclement weather
Accident	% of accidents thought to occur from driving in inclement weather

Data Files

Driving.xls

Links

Driving on Wet Roads. Jokes about Driving in Inclement Weather.

References

- Galski, T., Ehle, H. T, & Bradley, W. J. (1998). Estimates of driving abilities and skills in different conditions. American Journal of Occupational Therapy, 52, 268-275.
- Griffin, J., & Murdock, G. (1993, August). Wet weather driving. Consumers' Research Magazine, 76, 2.
- Taylor, G. W., & Quinn, H. (1991, January 14). An arctic winter rage. Maclean's, 104, 12-13.

14.13: Stroop Interference

Learning Objectives

- Strrop Interference demonstration

Research conducted by

Statistics Class

Case study prepared by

David Lane

Overview

Naming the ink color of color words can be difficult. For example, if asked to name the color of the word "blue" is difficult because the answer (red) conflicts with the word "blue." This interference is called "Stroop Interference" after the researcher who first discovered the phenomenon.

This case study is a classroom demonstration. Students in an introductory statistics class were each given three tasks. In the "words" task, students read the names of 60 color words written in black ink; in the "color" task, students named the colors of 60 rectangles; in the "interference" task, students named the ink color of 60 conflicting color words. The times to read the stimuli were recorded. There were 31 female and 16 male students.

Questions to Answer

Is naming conflicting color names faster or slower than naming color rectangles? Which is faster, naming color rectangles or reading color names? Are there gender differences?

Design Issues

This was not a well-controlled experiment since it was just a classroom demonstration. The order in which the students performed the tasks may not have been counterbalanced or randomized.

Descriptions of Variables

Table 14.13.1: Description of Variables

Variable	Description
Gender	1 for female, 2 for male
Words	Time in seconds to read 60 color words
Colors	Time in seconds to name 60 color rectangles
Interfer	Time in seconds to name colors of conflicting words

Data Files

Stroop.xls

Links

Full text of the above reference.

References

- Stroop, J.R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 28, 643-662.

14.14: TV Violence

Learning Objectives

- Does Television Viewing Encourage Aggression in Children?

Research conducted by

Mariana Fernandez, University of Houston-Downtown undergraduate

Case study prepared by

Nichole Rivera

Overview

How much television is too much for children? Television advocates espouse the educational benefits that children may reap from instructive programming. However, many researchers say that excess television watching may contribute to aggressive behavior in children. Young boys, in particular may be susceptible to this effect. What are the effects, if any, on children's behavior when television is used as a babysitter?

In a survey of University of Houston-Downtown students, parents reported their children's age, characteristic behavior, and television viewing habits. Convenience sampling was used to gather 30 subjects ($N = 30$).

Questions to Answer

Is there a relationship between hours of television watched and child's obedience? Will a child be more or less aggressive if he/she watches a lot of television?

Design Issues

This survey offered a very limited sample ($N = 30$), which was further hindered by reporting participants' filling out an individual survey for each individual child. This contributes to some lack of true variability in responses because participants tended to report similar behavior for each child. This may magnify errors associated with self-reported data. The sample would provide greater reliability if each participant reported on only one child's behavior.

The survey has broad questions which do not provide much context for reported behaviors. In some instances aggression may be positively rated, but this survey treats all aggression as a negative characteristic. In addition, the instrument itself measures largely nominal data, making in depth analysis difficult.

Descriptions of Variables

Table 14.14.1: Description of Variables

Variable	Description
TV hours	Total number of TV hours watched per day

Obedience	How obedient the child is 1 = very obedient, 5 = not obedient
Attitude	Attitude while playing with other children 1 = non-aggressive, 5 = very aggressive

Data Files

TV.xls

Links

TV Guide - Mighty Morphin' Power Rangers v. Teenage Mutant Ninja Turtles

References

- Boyatzis, Chris J. and Matillo Gina M. (1995). Effects of "The Mighty Morphin Power Rangers" on Children's Aggression with Peers. Child Study Journal, 25 (1), 45-57.
- Charlton, Davie. (2001). Monitoring Children's Behavior in Remote Community Before and Six Years After the Availability of Broadcast TV. North America Journal of Psychology, 3, 429-441.
- Huesmann, Rowell L., Moise-Titus, Jessica, Podolski, Cheryl-Lynn, Eron, Leonard D. (2003). Longitudinal Relations Between Children's Exposure to TV Violence and their Aggressive and Violent Behavior in Young Adulthood: 1977-1992. Developmental Psychology 39(2), 201-221.
- Troseth, Georgene L. (2003). TV Guide: Two-Year-Old Children Learn to Use Video as a Source of Information. Developmental Psychology 39 (1), 140-150.

14.15: Obesity and Bias

Learning Objectives

- Bias Against Associates of the Obese

Research conducted by

Mikki Hebl and Laura Mannix

Case study prepared by

Emily Zitek

Overview

Obesity is a major stigma in our society. People who are obese face a great deal of prejudice and discrimination. For example, Roehling (1999) showed that obese people experience a lot of discrimination in the workplace (e.g., they are less likely to be hired and get lower wages). We know that people who are obese are stigmatized, but what about people who are somehow associated with an obese person? Neuberg et al. (1994) found that friends of gay men and lesbians suffer from "stigma by association". Perhaps the negative effects of the obesity stigma can also spread to other people. This study seeks to examine how the stigma of obesity can spread to a job applicant of average weight.

As part of a larger study, participants had to rate how qualified a particular job applicant was. This applicant was sitting by a woman. The researchers manipulated the following two variables: the weight of the woman and the relationship between the woman and the applicant. The woman was either obese or of average weight. This woman was also portrayed as being the applicant's girlfriend or a woman simply waiting to participate in a different experiment.

Questions to Answer

Are male applicants who are seated next to an obese woman rated as less qualified for a job? Are applicants who are seated next to their girlfriend rated differently from applicants seated next to a woman with whom they do not have an intimate relationship? Finally, does the effect of the type of relationship differ depending on the weight of the woman?

Design Issues

This study only looked how at how an obese woman seated next to a male job applicant could affect qualification ratings. Future research could address other gender combinations.

Descriptions of Variables

Table 14.15.1: Description of Variables

Variable	Description
Weight	The weight of the woman sitting next to the job applicant 1 = obese, 2 = average weight
Relate	Type of relationship between the job application and the woman seated next to him 1 = girlfriend, 2 = acquaintance (waiting for another experiment)
Qualified	Larger numbers represent higher professional qualification ratings

Data Files

Weight2.xls

Links

The Obesity Society

References

- Hebl, M. R., & Mannix, L. M. (2003). The weight of obesity in evaluating others: A mere proximity effect. Personality and Social Psychology Bulletin, 29, 28-38.
- Neuberg, S. L., Smith, D. M., Hoffman, J. C., & Russell, F. J. (1994). When we observe stigmatized and "normal" individuals interacting: Stigma by association. Personality and Social Psychology Bulletin, 20, 196-209.
- Roehling, M. (1999). Weight-based discrimination in employment: Psychological and legal aspects. Personnel Psychology, 52, 969-1016.

14.16: Shaking and Stirring Martinis

Learning Objectives

- To test the difference between shaken and stirred martinis

Research conducted by

This is just made up data.

Case study prepared by

David Lane

Overview

This is an example to illustrate hypothesis testing and the binomial distribution. The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here is an example based on James Bond who insisted that Martinis should be shaken rather than stirred. In this hypothetical experiment to determine whether Mr. Bond could tell the difference between a shaken and a stirred martini, we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Mr. Bond was correct on $13/16$ trials.

Questions to Answer

Does Mr. Bond have the ability to tell the difference between a Martini that is shaken and one that is stirred?

Design Issues

This is only a made-up study.

Descriptions of Variables

Table $14.16.1$: Description of Variables

Variable	Description
Y	0 = incorrect, 1 = correct

Data Files

Martini.xls

Links

The Lady Tasting Tea

References

- Salsburg, D. (2002) The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. Owl Books

14.17: Adolescent Lifestyle Choices

Learning Objectives

- Adolescents and Healthy Lifestyle Choices

Research conducted by

Ka He, Ellen Kramer, Robert F. Houser, Virginia R. Chomitz, and Karen A. Hacker

Case study prepared by

Robert F. Houser, Alyssa Koomas, and Georgette Baghdady

Overview

Teen pregnancy, sexually transmitted disease, drug abuse, and suicide are some of the behaviorally-mediated negative health outcomes that can occur during adolescence. Identifying the characteristics of adolescents who are able to make healthy lifestyle choices is imperative toward understanding positive health behaviors in this age group. This information could be used to develop targeted interventions that support at-risk adolescents in making healthy lifestyle choices, and hopefully prevent such negative outcomes.

This study collected survey data from 1487 high school students in an urban Massachusetts community. The survey assessed health-related behaviors, stressful events, demographics, familial characteristics, perceptions of peer and parental support, and academic performance. In collaboration with community stakeholders and parents, the researchers selected six health-related behaviors and developed two sets of criteria to define positive health behaviors. One set used “strict” definitions, namely, not drinking alcohol in the last 30 days, no attempted suicide in the past 12 months, and no experience at all with tobacco, hard illegal drugs, marijuana, and sexual partners. The second set used “broad” definitions that allowed for mild use and safe experimentation (except for suicidal behavior). Students who adhered to all six health-related behaviors according to the “strict” definitions formed one subgroup for analysis, and those who reported behaviors in accordance with the “broad” definitions formed another subgroup. These two lifestyle subgroups were analyzed separately in relation to the personal and social-environmental factors assessed by the survey.

Questions to Answer

What personal and social-environmental characteristics are associated with adolescents who practice healthy lifestyle behaviors according to the “strict” definitions? How much more likely are adolescents with these characteristics to be practicing healthy behaviors than adolescents without these characteristics?

Design Issues

The results of this study may not be applicable to adolescents in non-urban schools, as the sample was drawn from a diverse, urban school. As well, the definitions that make up positive health behaviors may vary by region and social group. Adolescents self-reported their health-related behaviors and other information via the survey. Missing responses may have caused bias in the results.

Descriptions of Variables

Table 14.17.1: Description of Variables

Variable	Description
Healthy behaviors based on the “strict” definitions	Whether or not the adolescent practices all 6 health-related behaviors according to the “strict” definitions
Immigration status	Whether the adolescent was born in the US or is an immigrant

Stress score	An index from 0 to 14 assessing 14 possible stressful events in the adolescent's life, such as failing grades, moving, death in the family, divorce in the family, abuse, and violence
Stress index	Whether the adolescent's stress score is at or above the median stress score of 2, or below
Academic performance	The adolescent's average academic letter grade (A, B, C, D, F)

Links

1 in 3 Teens Text While Driving

2011 Youth Risk Behavior Surveillance Survey

References

- He, K., Kramer, E., Houser, R. F., Chomitz, V. R., Hacker, K. A. (2004). Defining and understanding healthy lifestyles choices for adolescents. Journal of Adolescent Health, 35, 26-33.

14.18: Chocolate and Body Weight

Learning Objectives

- To study chocolate's healthful metabolic mechanisms

Research conducted by

Beatrice A. Golomb, Sabrina Koperski, and Halbert L. White

Case study prepared by

Robert F. Houser and Georgette Baghdady

Overview

Recent research has brought to light the beneficial health effects of chocolate. Studies have linked chocolate with lower blood pressure, lower bad cholesterol, improved insulin sensitivity, and reductions in the risks of diabetes, heart disease, and stroke. The authors of this study hypothesized that chocolate's healthful metabolic mechanisms might also reduce fat deposition in spite of its high caloric content.

This study used the baseline data from a clinical study that examined noncardiac effects of cholesterol-lowering drugs in healthy adults. The baseline data included body mass index (BMI), chocolate consumption frequency, age, sex, physical activity frequency, depression, and some dietary variables. Chocolate consumption frequency was assessed with the question: "How many times a week do you consume chocolate?" Dietary intakes of total calories, fruits and vegetables, and saturated fat were assessed with a validated food frequency questionnaire. A food frequency questionnaire is a limited checklist of foods and beverages with a frequency response section for subjects to report how often each item was consumed over a specified period of time. Depression was measured with a validated scale related to mood. BMI is a measure of body fatness that is associated with many adverse health conditions.

Questions to Answer

What can we conclude from the researchers' findings that there is an association between consuming chocolate frequently and lower BMI? How do we interpret regression models?

Design Issues

The authors used baseline data from an unrelated clinical study examining noncardiac effects of cholesterol-lowering drugs. That clinical study included men ranging in age from 20 to 85 years, but only postmenopausal women. The results of the chocolate study cannot, therefore, be generalized to younger adult women. Except for BMI, the data for all of the study variables were "self-reported" by the subjects via questionnaires. The assessment of critical variables, such as chocolate consumption frequency and vigorous physical activity frequency, could differ when using different measurement tools. The study was cross-sectional in nature, precluding conclusions about causation.

Descriptions of Variables

Table 14.18.1: Description of Variables

VARIABLE	DESCRIPTION
BMI	Body mass index, calculated as: (weight in kilograms) / (height in meters)^2
Chocolate consumption frequency	Number of times per week a subject consumed chocolate
Calories	Overall calorie intake of a subject determined via food frequency questionnaire

Age	Range of 20 to 85 years, postmenopausal if female
Sex	68% male, 32% female
Activity	Number of times per 7-day period a subject engaged in vigorous physical activity for at least 20 minutes

Links

Golomb et al. article

Rose et al. article

What is body mass index (BMI)?

References

- Golomb, B. A., Koperski, S., White, H. L. (2012). Association between more frequent chocolate consumption and lower body mass index. Archives of Internal Medicine, 172, 519-521.
- Rose, N., Koperski, S., Golomb, B. A. (2010). Mood food: chocolate and depressive symptoms in a cross-sectional analysis. Archives of Internal Medicine, 170, 699-703.

14.19: Bedroom TV and Hispanic Children

Learning Objectives

- Study of overweight and obesity in Hispanic children

Research conducted by

Du Feng, Debra B. Reed, M. Christina Esperat, and Mitsue Uchida

Case study prepared by

Robert F. Houser, Alyssa Koomas, and Georgette Baghdady

Overview

The prevalence of overweight and obesity in children in the U.S. is a growing public health concern that disproportionately affects Hispanic youth. As noted by the authors, in 2005 to 2006, 15.5% of all U.S. children aged 2 to 19 years were overweight or obese, compared with 23.2% for boys and 18.5% for girls among Mexican-Americans in this age group. Past research has revealed diverse environmental and behavioral factors that may contribute to this disparity. For example, studies have shown that Hispanic children watch more television than white children.

This study examined TV viewing among 314 Hispanic children aged 5 to 9 years in West Texas and the possible effects of having a TV in the child's bedroom. Children's weights and heights were measured, body mass indexes (BMI) calculated, and sex- and age-adjusted BMI percentiles obtained. The 2000 CDC Growth Charts were used to assess whether or not a child was overweight or at risk for becoming overweight. Their parents completed a family survey assessing demographics, acculturation, parental support of physical activity, dietary practices, the presence of a TV in the participating child's bedroom, and the child's TV/DVD viewing time.

Questions to Answer

Do children with a TV in their bedroom spend more time watching TV/DVDs on a daily basis than children without a TV in their bedroom? Do children with a TV in their bedroom have less support from their parents for physical activity than children without a TV in their bedroom? What might account for missing responses to survey questions?

Design Issues

Except for BMI, the data for all of the study variables were "self-reported" by the parents. The study used a cross-sectional design, which cannot be relied upon to provide conclusive evidence of causal relationships.

Descriptions of Variables

Table $14.19.1$: Description of Variables

VARIABLE	DESCRIPTION
TVIB, No TVIB	Presence or absence of a TV in the participating child's bedroom
Daily TV/DVD time	Average number of hours the child spent watching TV and DVDs per day

Parental support of physical activity	Scale score calculated as the average of parent's responses to 8 survey items assessing the parent's support of physical activity for the child. Items rated on 4-point Likert scale (0 = never, 3 = always). Research has shown a significant positive relationship between parental support of physical activity and children's physical activity level
Daily fruit and vegetable intake	Average number of cups of fruits and vegetables (fresh, frozen, dried, canned, and 100% juice) consumed by the child per day
Daily sweetened beverages	Average number of ounces of soda, fruit drink, sports drink, tea, and lemonade consumed by the child per day

Links

New York Times article

Television and Children information guide

References

- Feng, D., Reed, D. B., Esperat, M. C., Uchida, M. (2011). Effects of TV in the bedroom on young Hispanic children. American Journal of Health Promotion, 25, 310-318.

14.20: Weight and Sleep Apnea

Learning Objectives

- Excess Body Weight and Sleep Apnea

Research conducted by

Kari Johansson, Erik Hemmingsson, Richard Harlid, Ylva Trolle Lagerros, Fredrik Granath, Stephan Rössner, and Martin Neovius

Statistical article authored by

Philip Sedgwick

Case study prepared by

Robert F. Houser and Georgette Baghdady

Overview

In his statistical article, "Standard deviation versus standard error," UK researcher Philip Sedgwick presents us with an interesting discussion of the proper use of standard deviation (SD) and standard error of the mean (SEM). He uses an example of a weight loss study of 63 obese men suffering from obstructive sleep apnea who were being treated with continuous positive airway pressure (CPAP). The weight loss program lasted one year. Outcome measures included change in body weight measured in kilograms (kg).

More than 60% of people experiencing obstructive sleep apnea are obese. CPAP therapy is the most common treatment. It uses a machine and mask to prevent the airway from collapsing, thus enabling a person to breathe more easily during sleep. Weight loss is an effective treatment for sleep apnea.

Questions to Answer

What is the proper use of the SD? What is the proper use of the SEM?

Design Issues

None for the Sedgwick article.

Descriptions of Variables

Table 14.20.1: Description of Variables

Variable	Description
Weight	Body weight at baseline in kg
Weight change	Change in body weight at one year from baseline in kg

Links

What Is Sleep Apnea?

t Table (two-tailed) for significance and calculation of confidence interval

Johansson et al. article

References

- Sedgwick, P. (2011). Standard deviation versus standard error. BMJ, 343, d8010.
- Johansson, K., Hemmingsson, E., Harlid, R., Lagerros, Y. T., Granath, F., Rössner, S., Neovius, M. (2011). Longer term effects of very low energy diet on obstructive sleep apnoea in cohort derived from randomised controlled trial: prospective observational follow-up study. BMJ, 342, d3017.

14.21: Misusing SEM

Learning Objectives

- Misusing Standard Error of the Mean (SEM)

Research conducted by

Peter Nagele

Case study prepared by

Robert F. Houser, Georgette Baghdady, and Jennifer E. Konick

Overview

Authors of published research articles often erroneously use the standard error of the mean to describe the variability of their study sample. Nagele demonstrated this misuse of the standard error of the mean as a descriptive statistic by manually searching four leading anesthesia journals in 2001.

Here are quotes on key points from Nagele's article and our notes:

"Descriptive statistics aim to describe a given study sample without regard to the entire population."

"If normally distributed, the study sample can be described entirely by two parameters: the mean and the standard deviation (SD)." However, a study sample variable is never exactly normally distributed. When a variable is close to normally distributed, the mean and median are quite similar. Therefore, the mean and SD would be sufficient.

"The SD represents the variability within the sample." It tells us about "the distribution of individual data points around the mean." The latter statement, however, is a generalization since the SD cannot tell us exactly where each data point lies relative to the mean.

"[I]nferential statistics generalize about a population on the basis of data from a sample of this population."

The standard error of the mean (SEM) "is used in inferential statistics to give an estimate of how the mean of the sample is related to the mean of the underlying population." It "informs us how precise our estimate of the [population] mean is."

Thus, "the SEM estimates the precision and uncertainty [with which] the study sample represents the underlying population."

The standard error of the mean is calculated by dividing the sample standard deviation by the square root of the sample size ($SEM = SD/\sqrt{n}$).

"[T]he SEM is always smaller than the SD." However, this is only true as long as the sample size is greater than 1.

"In general, the use of the SEM should be limited to inferential statistics [for which] the author explicitly wants to inform the reader about the precision of the study, and how well the sample truly represents the entire population [of interest]." A sample never truly represents the population.

Questions to Answer

How prevalent is the inappropriate use of the SEM in describing the variability of the study sample in research publications? What is the proper use of the SEM?

Design Issues

The author focused on four leading anesthesia journals in his field of expertise. The misapplication of the SEM in descriptive statistics can be found in professional journals of many, if not all, fields of research.

Descriptions of Variables

Table $14.21.1$: Description of Variables

Variable	Description

Incorrect use of SEM; total	Total frequency of misuse of SEM; expressed as number of articles and percent
Laboratory studies using SEM incorrectly	A subset of the above variable; expressed as number of articles and percent
Correct use of SD	Frequency of correct use of standard deviation; expressed as number of articles and percent

Data Files

Sem.xls

Links

Nagele article

References

- Nagele, P. (2003). Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. British Journal of Anaesthesia, 90, 514-516.
- Hassani, H., Ghodsi, M., Howell, G. (2010). A note on standard deviation and standard error. Teaching Mathematics and Its Applications, 29, 108-112.
- Altman, D. G., Bland, J. M. (2005). Standard deviations and standard errors. BMJ, 331, 903.

14.22: School Gardens and Vegetable Consumption

Learning Objectives

- School garden program benefits

Research conducted by

Michelle M. Ratcliffe, Kathleen A. Merrigan, Beatrice L. Rogers, and Jeanne P. Goldberg

Case study prepared by

Robert F. Houser and Georgette Baghdady

Overview

School garden programs are gaining popularity because of their numerous benefits for children: outdoor exercise, social skills, connecting with nature, environmental stewardship, active learning, experiential science education, higher academic achievement, and transformed attitudes and habits related to fruits and vegetables. By integrating the regular science class with gardening activities in which students plant, nurture, harvest, prepare, and consume produce grown in the schoolyard, studies are showing that garden-based learning can improve children's consumption of fruits and vegetables.

This study investigated the impact of participating in a school garden program on the ability to identify, willingness to taste, preference for, and consumption of vegetables. Subjects were 320 sixth-grade students aged 11 to 13 years at two intervention schools and one control school. At the intervention schools, garden-based learning activities were incorporated into the regular science class for a period of four months. The control school did not include a garden program as part of its science class. Two questionnaires – Garden Vegetable Frequency Questionnaire and taste test – assessed the outcome variables using vegetables typically grown in school gardens that were also ethnically and culturally appropriate for the study population. The Garden Vegetable Frequency Questionnaire assessed the types of vegetables consumed the day before as well as usual consumption frequency. The taste test involved tasting five raw vegetables (carrots, string beans, snow peas, broccoli, and Swiss chard). Both questionnaires were administered at the outset and end of the study. Change scores (posttest minus pretest) were compared between the garden (intervention) group and the control group.

Questions to Answer

Do hands-on school garden programs increase vegetable consumption in children? What are some of the potential sources of bias in research studies?

Design Issues

This study used a "quasi-experimental" design, which differs from an experiment in that the students were selected and assigned to the intervention group and control group by a method other than random assignment. With this type of design, there is a greater chance that the intervention and control groups might differ at the outset of the study in ways that could bias the results of the study. Since the study population was middle-school students living in low-income, urban communities, the results of the study cannot be generalized to other settings. The study did not measure the actual amounts of vegetables that students consumed, so no conclusions can be drawn about number or size of servings.

Descriptions of Variables

Tables* 14.22.1*: *Description of Variables*

Variable	Description
School garden program group	Garden (intervention) and Control groups: Whether or not a student experiences hands-on gardening activities at school

Consumption of vegetables at school	Assessed by the taste test, it measures whether or not a student ate each of five specific vegetables at school
Consumption of vegetables at home	Assessed by the taste test, it measures whether or not a student also ate each of the five specific vegetables at home

Links

Ratcliffe et al. article

The benefits of school gardens

First Lady Michelle Obama Hosts White House Garden Spring 2011 Planting

References

- Ratcliffe, M. M., Merrigan, K. A., Rogers, B. L., Goldberg, J. P. (2011). The effects of school garden experiences on middle school-aged students' knowledge, attitudes, and behaviors associated with vegetable consumption. Health Promotion Practice, 12, 36-43.

14.23: TV and Hypertension

Learning Objectives

- TV viewing time and adverse health

Research conducted by

Perrie E. Pardee, Gregory J. Norman, Robert H. Lustig, Daniel Preud'homme, and Jeffrey B. Schwimmer

Case study prepared by

Robert F. Houser and Andrew Kennedy

Overview

A strong, evidence-based association exists between TV viewing time and the risk of being obese in children and adolescents. Little or no research, however, has explored adverse health outcomes associated with TV viewing among obese children. This study aimed at identifying whether or not time spent watching TV is associated with hypertension (high blood pressure) in obese children.

Obese children aged 4 to 17 years were recruited and evaluated at three pediatric centers. Obesity was defined as a body mass index (BMI) greater than or equal to the 95th percentile for the child's age and gender.

Questions to Answer

Is TV watching associated with hypertension in obese children?

Design Issues

The study involved a cross-sectional design, which prevented the determination of possible causality among the associations found. There could be unmeasured factors that play a role in the association between TV viewing and hypertension.

Descriptions of Variables

Table $14.23.1$: Description of Variables

Variable	Description
Hypertension	Defined as a systolic and/or diastolic blood pressure greater than or equal to the 95th percentile for the child's age, gender, and height
Age	A child's age in years
BMI	A child's body mass index, calculated as: (weight in kilograms) / (height in meters)^2
Hours of TV/day	An estimate of a child's average daily time spent watching TV in hours

Links

Pardee et al. article

Luma et al. article

References

- Pardee, P. E., Norman, G. J., Lustig, R. H., Preud'homme, D., Schwimmer, J. B. (2007). Television viewing and hypertension in obese children. American Journal of Preventive Medicine, 33, 439-443.
- Luma, G. B., Spiotta, R. T. (2006). Hypertension in children and adolescents. American Family Physician, 73, 1558-1568.

14.24: Dietary Supplements

Learning Objectives

- Dietary supplements and health risk behaviors

Research conducted by

Wen-Bin Chiou, Chao-Chin Yang, and Chin-Sheng Wan

Case study prepared by

Robert F. Houser and Georgette Baghdady

Overview

Although the dietary-supplement market in the U.S. is enormous, there is no apparent association between the use of dietary supplements and improved public health. The researchers of this study explored this paradox under the hypothesis that taking dietary supplements triggers a phenomenon called the "licensing effect," namely, the tendency for positive choices to license subsequent self-indulgent, risky or unhealthful choices. The researchers hypothesized that supplement use confers "perceived health credentials," leading people to feel invulnerable to health hazards and thus more likely to engage in risky, health-related behaviors.

The study involved two experiments. In the first experiment, 82 participants were randomly assigned to either a vitamin-pill (multivitamin) group or control (placebo) group and were told the kind of pill they would be taking. However, only the control group was given correct information. In actuality, both groups received the placebo pill. After taking the pills, the participants completed a survey on leisure-time activities, rating the desirability of nine hedonic (pleasurable) activities, such as excessive drinking and wild parties, and nine exercise activities, such as yoga and running, on 7-point scales. The survey also included a general invulnerability scale to assess a participant's perceived invulnerability to harm and disease. After completing the survey, the participants were offered a free lunch, choosing freely between a buffet and a healthful, organic meal.

The second experiment involved different participants. The vitamin-pill (multivitamin) group again unknowingly took placebo pills. After completing a questionnaire that included the general invulnerability scale and reading a medical report on the health benefits of walking, the distance participants walked in one hour was measured with a pedometer.

Questions to Answer

Does taking dietary supplements disinhibit unhealthy behaviors, such as eating unhealthful meals? Is the study sufficiently powered to detect significant differences between males and females?

Design Issues

The research was conducted in Taiwan, where cultural attitudes and behaviors related to dietary supplements may differ from those in the U.S. It is possible that the results might not generalize to other countries, so more research is needed. Participants in Experiment 1 had a wide range in age, from 18 to 46 years, with a mean (SD) of 30.9 (7.8) years. It would be helpful to consider age in the analysis, especially if age is associated with invulnerability scores. Leisure-time activities and invulnerability were assessed only post-intervention; future studies should also measure these variables before the intervention to see if the two groups had similar scores at the start of the study. The general invulnerability scale used to assess perceived invulnerability to harm and disease has been validated only for adolescents.

Descriptions of Variables

Table $14.24.1$: Description of Variables

VARIABLE	DESCRIPTION
Experimental condition	Vitamin-pill (multivitamin) condition or Control (placebo) condition

Meal choice	Either a buffet meal or a healthful, organic meal
Gender	The sex of participants

Links

The licensing effect

No Significant Difference … Says Who?

References

- Chiou, WB, Yang, CC, Wan, CS. (2011). Ironic effects of dietary supplementation: Illusory invulnerability created by taking dietary supplements licenses health-risk behaviors. Psychological Science, 22, 1081-1086.
- Trout, A. T., Kaufmann, T. J., Kallmes, D. F. (2007). No significant difference … Says who? Editorial. American Journal of Neuroradiology, 28, 195-197.

14.25: Young People and Binge Drinking

Learning Objectives

- Binge drinking and serious public health problems

Research conducted by

Richard O. de Visser and Julian D. Birch

Case study prepared by

Robert F. Houser and Georgette Baghdady

Overview

Binge drinking is a serious public health problem bringing harm to both the individual and society. It compromises a person's health, increasing the risk of many diseases, injury, and death. It also results in a greater incidence of motor vehicle crashes, violence, the spread of sexually-transmitted diseases, and unintended pregnancies. Binge drinking is prevalent among both young and older adults, men and women, and high and low income levels. Governments have formulated guidelines for moderate or sensible drinking levels. The government of the United Kingdom (UK) issued guidelines for sensible drinking as $2-3$ alcohol units per day for women and $3-4$ units per day for men, an alcohol unit being 10 milliliters of ethanol. A binge drinking episode is when a person drinks above double the recommended daily guidelines in a short period of time.

Questions to Answer

What can we learn about the binge drinking patterns of university students in England? Do the bingers and non-bingers differ in their knowledge of the sensible drinking guidelines issued by the UK government?

Design Issues

The university students in the sample "self-selected" to participate in the study by responding to recruiting efforts made via email messages and requests in lectures.

Descriptions of Variables

Table $14.25.1$: Description of Variables

Variable	Description
Sex	Female or male
mo_binge_n	Number of times the university students did binge drinking in the last month (using sex-specific definitions)
modrunk	Number of times the university students drank in the last month
wk_unit_prop	Familiarity with alcohol unit-based guidelines (measured on a 5-point scale)
k_unit_sum	Knowledge of alcohol unit-based guidelines (score out of 7)
u_fam	Familiarity with alcohol unit-based guidelines (measured on a 5-point scale)

Data Files

Binge.xls

Links

de Visser et al. article

References

- de Visser, R. O., Birch, J. D. (2012). My cup runneth over:Young people's lack of knowledge of low-risk drinking guidelines. Drug and Alcohol Review, 31, 206-212.

14.26: Sugar Consumption in the US Diet

Learning Objectives

- Sugar Consumption in the US Diet between 1822 and 2005

Research conducted by

Stephan Guyenet and Jeremy Landen

Case study prepared by

Robert F. Houser and Georgette Baghdady

Overview

Sugar has many forms: cane sugar, beet sugar, honey, molasses, fruit juice concentrate, glucose, sucrose, fructose, high-fructose corn syrup, maple syrup, brown rice syrup, barley malt syrup, agave nectar, to list a few. High-fructose corn syrup, in particular, was introduced into the US food industry in the early $1970s$ and has become ubiquitous in processed foods and soft drinks. Many of the added sugars in packaged foods and beverages could be considered "hidden sugar" because, if we do not examine the ingredients list on food labels or know sugar's many aliases, we are most likely unaware of how much sugar we consume each day.

To explore sugar consumption trends in the US, researchers Stephan Guyenet and Jeremy Landen compiled data on caloric sweetener sales spanning 184 years. They extracted annual caloric sweetener sales per capita for 1822 to 1908 from US Department of Commerce and Labor reports, and for 1909 to 2005 from the US Department of Agriculture (USDA) web site. The researchers adjusted the sales data for post-production losses using the USDA's $1970-2005$ loss estimate of 28.8 percent to obtain reasonable estimates of annual per capita consumption of added sugars. Post-production losses of a food commodity occur at the retail, foodservice and consumer levels from, for example, spoilage, pests, cooking losses and plate waste.

Guyenet presents a striking graph and regression analysis of sugar consumption in the US from 1822 to 2005 in a blog to promote awareness and discussion.

Questions to Answer

Do different time periods between 1822 and 2005 reveal different trends in sugar consumption in the US diet? Can a regression graph be used to make predictions outside the range of the study data?

Design Issues

The data represent added caloric sugars such as cane sugar, high-fructose corn syrup and maple syrup, not naturally occurring sugars such as those in fruits and vegetables. Thus the data do not represent total sugar consumption. The data are not direct measures of consumption, but rather estimates derived from sales figures by adjusting for losses before consumption. The adjustment, applied across all years, is based on the USDA loss estimate from $1970-2005$, which may or may not underestimate sugar consumption in earlier time periods.

Descriptions of Variables

Table $14.26.1$: Description of Variables

Variable	Description
year	All years from 1822 to 2005
sugar_consum	Estimated consumption of added sugars in the US diet in pounds per year per person

Data Files

Sugar.xls

Links

By 2606, the US Diet will be 100 Percent Sugar, a blog by Stephan Guyenet

How to Spot Added Sugar on Food Labels

Dietary Sugars Intake and Cardiovascular Health: A Scientific Statement From the American Heart Association

Sugar: The Bitter Truth, a lecture by Robert H. Lustig

60 Minutes: Is Sugar Toxic?

References

- Johnson, R. K., Appel, L. J., Brands, M., Howard, B. V., Lefevre, M., Lustig, R. H., Sacks, F., Steffen, L. M., Wylie-Rosett, J. (2009). Dietary sugars intake and cardiovascular health: A scientific statement from the American Heart Association. Circulation, 120, 1011-1020.

14.27: Nutrition Information Sources and Older Adults

Learning Objectives

- Better educated people and information sources

Research conducted by

Diane L. McKay, Robert F. Houser, Jeffrey B. Blumberg, and Jeanne P. Goldberg

Case study prepared by

Robert F. Houser, Alyssa Koomas, Georgette Baghdady, and Jennifer E. Konick

Overview

Various socioeconomic factors, such as occupation, income, race, and education level, are associated with health outcomes. Prominent among them, education level has proved to be a strong predictor of diet quality, health behavior patterns, and disease risk. Studies have found that better-educated people have healthier diets than those with less education, leading some researchers to hypothesize that better-educated people may obtain nutrition information from more reliable sources than less-educated people.

This study examined that hypothesis among a sample of 176 adults aged 50 years or older. The participants completed a survey which asked whether or not they primarily relied upon each of the following sources for information about nutrition: doctors, other medical professionals, newspapers, magazines, television, radio, friends, relatives, and neighbors. Analysis involved comparing the sources by education level. Older adults are highly vulnerable to diet-related disease. Knowing which sources they rely on can enable nutrition educators and professionals to target those sources with high-quality nutrition messages, tailored to the needs and education level of the older-adult audience.

Questions to Answer

What sources of nutrition information do older adults rely on? Do these sources differ according to the educational attainment and gender of the adults? Are these sources of nutrition information related to dietary practices, such as taking supplements?

Design Issues

Given that the sample was drawn only from the New England area and that 93% were Caucasian, the results of this study should not be generalized to older adults in other regions or racial and ethnic groups. The Internet as a source of nutrition information was not included in the survey; it is likely a primary source among today's older adults.

Descriptions of Variables

Table $14.27.1$: Description of Variables

Variable	Description
coll4yrplus	Highest level of education completed: 0 = "< 4 years of college" (i.e., secondary school, high school, vocational school, community or junior college) 1 = "≥ 4 years of college" (i.e., four-year college, graduate or professional school)
gender	1 = female, 2 = male
doctor	Is your doctor a primary source of information about nutrition? 1 = yes, 2 = no

magazine	Are magazines a primary source of information about nutrition? 1 = yes, 2 = no
tv	Is TV a primary source of information about nutrition? 1 = yes, 2 = no
friends	Are friends a primary source of information about nutrition? 1 = yes, 2 = no
supps	Are you taking any dietary supplements? 1 = yes, 2 = no

Data Files

Nutrition_information.xls

Links

Nutrition Information For You

Evaluating Nutrition Information (see pages 36-43)

Nutrition Accuracy in Popular Magazines

References

- McKay, D. L., Houser, R. F., Blumberg, J. B., Goldberg, J. P. (2006). Nutrition information sources vary with education level in a population of older adults. Journal of the American Dietetic Association, 106, 1108-1111.

14.28: Mind Set - Exercise and the Placebo Effect

Learning Objectives

- The "placebo" effect

Research conducted by

Alia J. Crum and Ellen J. Langer

Case study prepared by

Robert F. Houser and Alyssa Koomas

Overview

The "placebo effect" is an effect that cannot be attributed to a drug or remedy, but rather to a change in a person's mind-set or perception. The placebo effect is widely accepted in clinical trials and its effects may shock you. For instance, one study found that subjects developed real rashes after being exposed to fake poison ivy (Blakeslee, 1998)! This study examined the placebo effect with relation to physical activity and health. Could becoming aware of how much you exercise result in weight loss even if you didn't make any changes to your diet or exercise routine?

The subjects were 84 female maids of ages 19 to 65 years at seven hotels. They were told the purpose of the study was to improve the health and happiness of hotel maids. According to the authors, "[e]ach of seven hotels was randomly assigned to one of two conditions: informed or control" (page 166). "Four hotels were assigned to the informed condition, and three were assigned to the control condition" (pages 166 – 167). Each subject filled out a questionnaire asking about her perceived amount of exercise during and outside of work. Physiological measurements were taken for weight, body mass index, body-fat percentage, waist-to-hip ratio, and blood pressure. The maids in the informed condition were then given an oral presentation and handouts explaining how their work as hotel maids is good exercise, so good in fact that it meets or exceeds the Surgeon General's recommendations for physical activity. The maids in the control condition were not given this information. After four weeks, the researchers re-administered the questionnaire and took follow-up physiological measurements.

Questions to Answer

Does the placebo effect play a role in the health benefits of exercise? If we alter a person's perception of the exercise she performs, does it result in weight loss?

Design Issues

Instead of assigning individual maids randomly to either the informed or control condition, all of the maids in the same hotel were assigned to the same condition. This was done in an effort to prevent information contamination. This type of study design is known as a "cluster randomized trial," and calls for advanced statistical practices that we will not worry about in this case study.

Simple random sampling with a sufficient number of subjects randomly assigned to intervention and control groups ideally leads to intervention and control groups that are similar with respect to many demographic characteristics. Simple random sampling of individuals and random assignment of individuals to conditions were not used in this study. The authors of this study pointed out that "[s]ubjects in the informed group were significantly younger than subjects in the control group." Consequently, they attempted to control for age differences in their statistical analysis.

The questionnaire asked about self-reported levels of exercise and dietary intake. Future research should use more rigorous methods to assess physical activity and diet.

Descriptions of Variables

Table 14.28.1: Description of Variables

VARIABLE	DESCRIPTION
cond	Condition: Either Informed or Control

age	Age in years
ex1	Perceived amount of exercise at Time 1 (On a scale from 0 to 10 with 0 = “none” and 10 = “a great deal”)
ex2	Perceived amount of exercise at Time 2 (On a scale from 0 to 10 with 0 = “none” and 10 = “a great deal”)
wt1	Weight in pounds at Time 1
wt2	Weight in pounds at Time 2
aex	Change score for exercise equal to the perceived amount of exercise at Time 2 minus the perceived amount of exercise at Time 1
awt	Weight change equal to the weight at Time 2 minus the weight at Time 1

Data Files

Mindset.xls

Links

Crum et al. article

New York Times article

References

- Crum, A. J., Langer, E. J. (2007). Mind-set matters: Exercise and the placebo effect. Psychological Science, 18, 165-171.

14.29: Predicting Present and Future Affect

Learning Objectives

- To explore the phenomenon of future anhedonia

Research conducted by

Karim S. Kassam, Daniel T. Gilbert, Andrew Boston, and Timothy D. Wilson

Case study prepared by

Robert F. Houser and Georgette Baghdady

Overview

In Aesop's fable, "The Ant and the Grasshopper," an ant toils all summer to gather food for the winter while a grasshopper sunbathes and enjoys the present abundance of food without concern for the upcoming winter. Consequently when winter arrives, the grasshopper despairs that it has no food. The moral of the fable is that it is best to prepare for the days of necessity. Clearly the grasshopper failed to predict accurately how it would feel in the winter while it sunbathed with a full belly in the summer.

The authors of this study explored the intriguing phenomenon of future anhedonia and its relation to the concept of time discounting in order to understand people's predictions about how they might feel when a future event happens. Time discounting occurs when people put less value on future events than present events. Future anhedonia refers to people's mistaken belief that a future event would elicit a less intense affective reaction than if the same event happened in the present. In six experiments, the authors asked participants to predict how happy they would feel both in the present and in the future upon receiving either 20 dollars outright or 25 dollars in the form of a Starbucks coffeehouse gift card. The difference between the scores of present and future happiness is a measure of future anhedonia.

Questions to Answer

Do people expect their affective reactions to an event to be less intense in the future than in the present?

Design Issues

The monetary amount ($\$20, \$25$) may not have been enough to psychologically engage a large number of participants. The wide age range in **Experiment 1b** of 15 to 72 years is unusual for a psychological study. Also in **Experiment 1b**, several participants reported that they would pay $\$25$ for a $\$25$-gift card that Starbucks was considering selling at a discounted price, which might indicate that they did not fully understand the question.

Descriptions of Variables

Table $14.29.1$: Description of Variables

VARIABLE	DESCRIPTION
Gender	The sex of a participant
Happiness score	A participant's estimate of his/her affective reaction to an event using a 9-point scale with endpoints 1 = "not at all happy" and 9 = "extremely happy"
diff_happy	A difference score equal to a participant's predicted present happiness score for a present event minus his/her predicted future happiness score for the same event in the future. A positive difference indicates future anhedonia.

diff_money	The difference in the maximum amount of money that a participant predicted as his/her willingness to pay in the present minus the predicted amount he/she would pay at a future time for a $25 Starbucks coffeehouse gift card. A positive difference indicates future anhedonia.
cond	Condition: Whether an event was expected or unexpected
today	A participant's predicted present happiness score for a present event
future	A participant's predicted future happiness score for a future event

Data Files

Predicting.xls

Links

Kassam et al. article

Aesop's Fable: The Ant and the Grasshopper

Prospection: Experiencing the Future

References

- Kassam, K. S., Gilbert, D. T., Boston, A., Wilson, T. D. (2008). Future anhedonia and time discounting. Journal of Experimental Social Psychology, 44, 1533-1537.
- Gilbert, D. T., Wilson, T. D. (2007). Prospection: Experiencing the future. Science, 317, 1351-1354

14.30: Exercise and Memory

Learning Objectives

- To study the benefits of exercise on memory

Research conducted by

M. E. Hopkins, F. C. Davis, M. R. Van Tieghem, P. J. Whalen, and D. J. Bucci

Case study prepared by

Robert F. Houser and Georgette Baghdady

Overview

Physical exercise has many beneficial effects on physiological processes, including those that affect cognition and memory. Exercise increases brain-derived neurotrophic factor (**BDNF**), which is a protein found in the learning and memory centers of the brain where it supports nerve cell survival and the growth of new neurons and neuronal connections. A polymorphism of **BDNF** (a variant genotype) alters the release of **BDNF** during exercise. The researchers of this study sought to compare the effects of a single bout of exercise versus a 4-week exercise regimen on cognition and memory and to determine if **BDNF** genotype influences the intensity of those effects of exercise.

Questions to Answer

How do regular exercise and/or an acute bout of exercise affect cognitive memory? Does type of **BDNF** genotype (Val/Val or Met carrier) mediate the effect of exercise on memory? How do we calculate a one-way ANOVA by hand and how do different post-hoc tests compare?

Design Issues

The group sample sizes are small, perhaps limiting the power to detect significant differences between the four exercise/control groups.

Descriptions of Variables

Table 14.30.1: Description of Variables

Variable	Description
Group	0W-: sedentary group 0W+: sedentary group with one bout of exercise at least 2 hours before Visit 2 4W-: regularly exercising group 4W+: regularly exercising group with a bout of exercise at least 2 hours before Visit 2
Accuracy	The percentage of objects each group accurately identified as old or new when performing the novel object recognition task during each study visit
Difference score	Accuracy achieved by the subject in the novel object recognition task during Visit 2 minus accuracy during Visit 1, in percent

BDNF genotype	Whether a subject's BDNF genotype is Val/Val or Met carrier (Val/Met and Met/Met)

Links

How Exercise Affects the Brain: Age and Genetics Play a Role

BDNF

References

- Hopkins, M. E., Davis, F. C., Van Tieghem, M. R., Whalen, P. J., Bucci, D. J. (2012). Differential effects of acute and regular physical exercise on cognition and affect. Neuroscience, 215, 59-68.

14.31: Parental Recognition of Child Obesity

Learning Objectives

- To study the parents' perception of their children's weight status

Research conducted by

Debra Etelson, Donald A. Brand, Patricia A. Patrick, and Anushree Shirali

Case study prepared by

Robert F. Houser and Georgette Baghdady

Overview

With increasing public awareness of child obesity as a major public health problem, studies are showing that it has not translated into an increased awareness of obesity in one's own child. Dietary patterns and weight status in childhood tend to carry into adolescence and adulthood, promoting the onset of chronic and other diseases. A key ingredient for combating childhood obesity is parental involvement and commitment. However, this is predicated on whether or not parents can recognize overweight and obesity in their children.

This study examined parents' perceptions of their children's weight status, their understanding of the health risks of obesity relative to other conditions they may perceive as health risks, and their knowledge of some healthy eating practices. Children's actual weight status was expressed as their body mass index (**BMI**) percentile, as determined by the CDC growth charts based on age and sex. According to the CDC growth charts for children, a child with a **BMI** percentile less than the 5th percentile is underweight; from the 5th to less than the 85th, a child is at a healthy weight; from the 85th to less than the 95th percentile, a child is overweight; and a **BMI** percentile equal to or greater than the 95th percentile, a child is considered to be obese.

A visual analog scale was used to measure parents' perceptions of their child's weight. The visual analog scale consisted simply of a 10-cm straight line anchored at the left end by the label "extremely underweight" and at the right end by the label "extremely overweight." A parent placed a mark along the line to indicate where they perceived their child's weight to be. The researchers interpreted the marks as percentiles in their analysis.

Questions to Answer

Do parents recognize when their children are overweight or obese? Do parents who make incorrect judgments about healthy food practices also make incorrect judgments about their child's weight status?

Design Issues

This study defines a parent's perception of their child's **BMI** percentile as "accurate" if their score on a visual analog scale fell within 30 points of the child's true **BMI** percentile. This wide range defining accuracy potentially allows for misclassification of a child's weight status among normal, overweight, and obese categories. For example, a parent who perceives their child's weight status as being at the 80th percentile, i.e., in the normal range, when in reality the child is obese with a **BMI** percentile of 98, the parent's assessment would be considered accurate by the operational definition used in this study. The authors explain that they chose this definition to give parents as much leeway as possible in assessing their child's weight on the visual analog scale.

Descriptions of Variables

Table 14.31.1: Description of Variables

Variable	Description
Sex	The sex of the participating parent's child

Overwt_Obese	Whether or not a child's body mass index (BMI) is equal to or greater than the 85th percentile for the child's age and sex, which means that the child is either overweight (85th to less than 95th percentile) or obese (95th percentile or above)
PA_overwt	Parental attitude expressing level of concern if their child were overweight, measured on a 4-point Likert Scale. In data analysis, the four categories were condensed into two categories: 0 = "not at all" or "a little" concerned 1 = "quite" or "extremely" concerned
PA-TV	Parental attitude expressing level of concern if their child watched >20 hours of TV per week, measured on a 4-point Likert Scale. In data analysis, the four categories were condensed into two categories: 0 = "not at all" or "a little" concerned 1 = "quite" or "extremely" concerned
Accurate	Whether or not the parent's perception of their child's weight status was accurate. Parent's perception was considered accurate if the BMI percentile it corresponded to fell within 30 points of the child's actual BMI percentile
Juice_boxes	The amount of juice that a parent thinks is healthy for their child to drink each day (a juice box contains eight ounces). We condensed the original four response categories into two categories: 0 = "1 or 2 juice boxes per day" 1 = "3 to 8 juice boxes per day"
Fast_food_meals	How often a parent feels it is okay to eat at fast-food restaurants. We condensed the original four response categories into two categories: 0 = "once a month"

Links

Etelson et al. article

BMI percentiles for children

References

- Debra Etelson, D., Brand, D. A., Patrick, P. A., Shirali, A. (2003). Childhood obesity: Do parents recognize this health risk? Obesity Research, 11, 1362-1368

14.32: Educational Attainment and Racial, Ethnic, and Gender Disparity

Learning Objectives

- To study large disparities in educational attainment among various racial and ethnic groups

Research conducted by

United States Census Bureau

Case study prepared by

Robert F. Houser, Georgette Baghdady, and Jennifer E. Konick

Overview

The U.S. Census Bureau defines educational attainment as the highest level of education that a person has completed. Large disparities in educational attainment continue to exist among racial and ethnic groups. The gender gap in educational attainment, however, has been undergoing a dramatic social shift in recent decades. In Table 14.32.1below, the U.S. Census Bureau tabulated these trends among Whites, Blacks, Asians and Pacific Islanders, and Hispanics between 1970 and 2010. This case study focuses only on college graduates. The data for "College graduate or more" represent the percentage of adults aged 25 years and older that obtained a degree from regular four-year colleges and universities and graduate or professional schools in each racial and ethnic group.

The U.S. Census Bureau defines the racial and ethnic categories in the following manner:

- “White” refers to persons having origins in any of the original peoples of Europe, the Middle East, or North Africa.
- “Black” refers to persons having origins in any of the Black racial groups of Africa.
- “Asian” refers to persons having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent.
- “Pacific Islander” refers to persons having origins in any of the original peoples of the Pacific Islands, such as Hawaii, Guam, Samoa, and Tonga.
- “Hispanic” refers to an ethnic group comprised of persons of any race who are of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin.

Table 230. Educational Attainment by Race, Hispanic Origin, and Sex: 1970 to 2010

[In percent. See Table 229 for headnote and totals for both sexes]

Year	All races [1]		White [2]		Black [2]		Asian and Pacific Islander [2]		Hispanic [3]	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
HIGH SCHOOL GRADUATE OR MORE [4]										
1970	51.9	52.8	54.0	55.0	30.1	32.5	61.3	63.1	37.9	34.2
1980	67.3	65.8	69.6	68.1	50.8	51.5	78.8	71.4	45.4	42.7
1990	77.7	77.5	79.1	79.0	65.8	66.5	84.0	77.2	50.3	51.3
1995	81.7	81.6	83.0	83.0	73.4	74.1	(NA)	(NA)	52.9	53.8
2000	84.2	84.0	84.8	85.0	78.7	78.3	88.2	83.4	56.6	57.5
2005	84.9	85.5	85.2	86.2	81.0	81.2	[5] 90.4	[5] 85.2	57.9	59.1
2007	85.0	86.4	85.3	87.1	81.9	82.6	89.8	85.9	58.2	62.5
2008	85.9	87.2	86.3	87.8	81.8	84.0	90.8	86.9	60.9	63.7
2009	86.2	87.1	86.5	87.7	84.0	84.1	90.4	86.2	60.6	63.3
2010	86.6	87.6	86.9	88.2	83.6	84.6	91.2	87.0	61.4	64.4
COLLEGE GRADUATE OR MORE [4]										
1970	13.5	8.1	14.4	8.4	4.2	4.6	23.5	17.3	7.8	4.3
1980	20.1	12.8	21.3	13.3	8.4	8.3	39.8	27.0	9.4	6.0
1990	24.4	18.4	25.3	19.0	11.9	10.8	44.9	35.4	9.8	8.7
1995	26.0	20.2	27.2	21.0	13.6	12.9	(NA)	(NA)	10.1	8.4
2000	27.8	23.6	28.5	23.9	16.3	16.7	47.6	40.7	10.7	10.6
2005	28.9	26.5	29.4	26.8	16.0	18.8	[5] 54.0	[5] 46.8	11.8	12.1
2007	29.5	28.0	29.9	28.3	18.0	19.0	55.2	49.3	11.8	13.7
2008	30.1	28.8	30.5	29.1	18.7	20.4	55.8	49.8	12.6	14.1
2009	30.1	29.1	30.6	29.3	17.8	20.6	55.7	49.3	12.5	14.0
2010	30.3	29.6	30.8	29.9	17.7	21.4	55.6	49.5	12.9	14.9

NA Not available. [1] Includes other races not shown separately. [2] Beginning 2005, for persons who selected this race group only. See footnote 2, Table 229. [3] Persons of Hispanic origin may be any race. [4] Through 1990, completed 4 years of high school or more and 4 years of college or more. [5] Starting in 2005, data are for Asians only, excludes Pacific Islanders.

Source: U.S. Census Bureau, U.S. Census of Population, 1970 and 1980, Vol. 1; Current Population Reports P20-550, and earlier reports; and "Educational Attainment," <http://www.census.gov/population/www/socdemo/educ-attn.html>.

Figure 14.32.1: Educational Attainment by Race, Hispanic Origin and Sex

Educational attainment is strongly associated with future employment, income, and health status.

Questions to Answer

How has the percentage of college graduates changed over time between 1970 and 2010 among the racial and ethnic groups and between the genders within each group? How might we illustrate these changes graphically?

Design Issues

Beginning with the 2000 U.S. Census, respondents were given the option of selecting more than one race category to indicate their racial identities. Therefore, data on race from 2000 and beyond are not directly comparable with earlier censuses. The data in Table 14.32.1represent persons who selected only one race category and exclude persons who selected more than one race.

In the 2005 U.S. Census and beyond, the "Asian and Pacific Islander" category was split into two separate categories, "Asian" and "Native Hawaiian or Other Pacific Islander." There were several reasons for the split. The combined category was not a homogeneous group because it put together peoples with few social or cultural similarities and who are dissimilar on important demographic characteristics. For example, in 1990, about 11 percent of Pacific Islanders aged 25 years and older obtained a bachelor's degree compared with about 40 percent of Asians. Since Pacific Islanders are numerically a smaller group than Asians (in 2010, there were about a half million Pacific Islanders versus about 14.6 million Asians), not including them in the data of Table 14.32.1starting in 2005 biases the percentage of college graduates upwards somewhat, but not strongly.

Descriptions of Variables

Table 14.32.1: Description of Variables

Variable	Description
College graduate or more	Obtained a degree from regular four-year colleges and universities and graduate or professional schools
Year	Decade years from 1970 to 2010
White_M White_F	Percentage of college graduates in U.S. subpopulation of White males aged 25 years and over; likewise for White females
Black_M Black_F	Percentage of college graduates in U.S. subpopulation of Black males aged 25 years and over; likewise for Black females
AsnPac_M AsnPac_F	Percentage of college graduates in U.S. subpopulation of Asian and Pacific Islander males aged 25 years and over; likewise for Asian and Pacific Islander females
Hispan_M Hispan_F	Percentage of college graduates in U.S. subpopulation of Hispanic males aged 25 years and over; likewise for Hispanic females

Data Files

Educational_attainment.xls

Links

Overview of Race and Hispanic Origin: 2010

Latinos and Education: Explaining the Attainment Gap

Why Do Women Outnumber Men in College?

References

- U.S. Census Bureau, Statistical Abstract of the United States: 2012. Section 4. Education, 143-151

- Telfair, J., Shelton, T. L. (2012). Educational attainment as a social determinant of health. North Carolina Medical Journal, 73(5), 358-365

CHAPTER OVERVIEW

15: Calculators

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

15.1: Analysis Lab

Learning Objectives

- To study the Analysis Lab calculator

General Instructions

Analysis Lab that can perform basic descriptive and inferential statistics. Descriptive statistics include the graphic methods: box plots, histograms, stem and leaf displays, and normal quantile plots. Inferential statistics include independent-group t tests, Chi Square tests, tests in simple regression/correlation, and Analysis of Variance.

Reading/Entering Data

There are currently two "libraries" of datasets. The default library is "RVLS_data." You can change the library by using the pop-up menu. Just under the pop-up menu for libraries is a pop-up menu of datasets in the library. Choose the dataset you want to analyze. A description of the dataset will be shown on the right side of the display. The data from the case studies will be included in the near future.

To enter your own data, click the button labeled "Enter/Edit User Data." A window will open with an area for you to enter or paste in your data. You may get a warning that you must use the keyboard shortcut to paste because your system may not allow Java programs to read the clipboard. Pasting formatted text will not work. If you have trouble pasting in your data, try pasting them as unformatted text into Word, copying, and then pasting into Analysis Lab.

The first line should contain the names of the variables (separated by spaces or tabs). The remaining lines should contain the data themselves. Missing data cannot be handled so all observations must have valid data for all variables. All variables must be numeric. If one of your variables is to be used as a "Grouping" or "Classification" variable, then values of the grouping variable must be integers ranging from one to the total number of groups. Use grouping variables so that you can do a separate analysis for each level of the variable or to use the variable as an independent variable in an analysis of variance. Once you have entered your data, click on the "Accept data" button. The data will be temporarily saved so that if you click the "Enter/Edit User Data" button again they will be shown

Choosing Variables

To analyze a variable, select it in the "Y" pop-up menu and then click on the type of analysis you wish to perform. To see the relationship between two variables, select one in the "X" menu and the other in the "Y" menu. Then click on the "Correlation/regression" button. Specify a grouping variable to do an analysis separately for each group of observations. You also specify a grouping variable to conduct an analysis of variance or do a t test.

Copying/Printing

Analysis Lab does not have any copying or printing capabilities. However, you can copy and(optionally) print by doing a screen capture. If you are using a Macintosh with OSX, press "4" while holding down the command (Apple), shift, and control keys. You will be given the opportunity to select an area of the screen to be copied to the clipboard.

If you are using Windows, make sure the window you want to copy is selected. Then, hold down the ALT key and press the Print Screen Key key. The window will be copied to the clipboard for pasting into another application such as a word processor.

Problems

On occasion, some browsers do not refresh the display. If you get a blank window, try resizing it slightly.

ANOVA Instructions

The following information is a guide to doing Analysis of Variance (ANOVA) with Analysis Lab. Analysis Lab was designed to allow the most common types of designs to be analyzed with a minimum of effort.

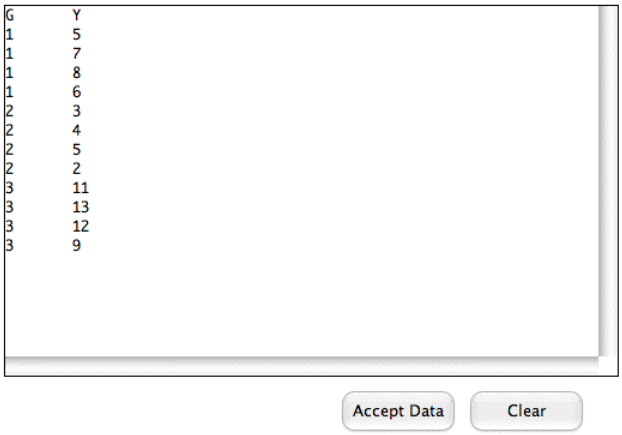
One Between-Subjects Variable

This analysis is used when there are two or more independent groups of subjects and one score per subject.

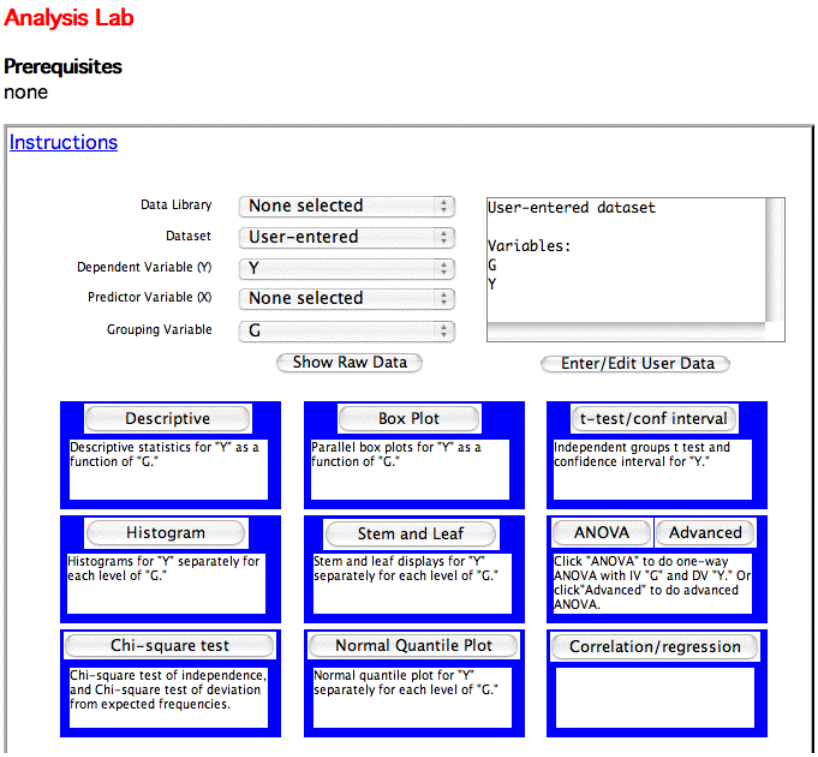
1. The data should be arranged so that there is one variable that indicates group membership and a second variable that contains the dependent variable. The variable indicating group membership must use consecutive integers (e.g., $1, 2, 3$) as group indicators. The data below have two variables; G for group membership and Y for the dependent variable. There are three groups and four observations (subjects) per group.

G	Y
1	5
1	7
1	8
1	6
2	3
2	4
2	5
2	2
3	11
3	13
3	12
3	9

2. Paste the data into analysis lab.
   a. Copy your data.
   b. Click the Enter/Edit data button.
   c. Paste the data using keyboard shortcuts (Control-V for Windows, CMD-V for Macs). Then click "Accept Data."



Choose Y as the dependent variable and G as the grouping variable as shown. Then click the ANOVA button. Note that the description says what analysis will be done. IV stands for "Independent Variable" and DV stands for "Dependent Variable."



d. The ANOVA summary table will be presented.

Source	df	SSQ	MS	F	p
G	2	122.1667	61.0833	29.3200	0.0001
Error	9	18.7500	2.0833		
Total	11	140.9167			

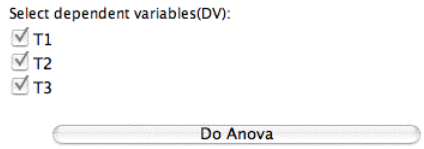
One Within-Subjects Variable

Use this design when you have one group of observations (subjects) with two or more scores per observation.

1. Format your data so that each row contains all the data for one observation. Each score is a separate variable. For the following example, there are five observations with three scores per observation. The variable names are $T1$, $T2$, and $T3$.

T1	T2	T3
2	3	4
3	5	4
2	4	6
1	3	3
6	7	9

2. Paste the data into analysis lab (see the instructions for One-Way Between-Subjects).
3. Click the "Advanced" button next to the ANOVA" button.
4. Select the dependent variables you want to analyze. and click the "Do Anova" button. Here the variables $T1$, and $T2$, and $T3$ are chosen.



5. The Summary Table is presented. The effect DV stands for "Dependent Variable." It is a test of the within-subjects variable.

Source	df	SSQ	MS	F	p
Subjects	4	44.4000	11.1000		
DV	2	14.9333	7.4667	13.5758	0.0027
Error	8	4.4000	0.5500		
Total	14	63.7333			

Two or More Between-Subject Variables

Use this design when you have more than one between-subject variable and only one score per subject. In the following example, the variables are A and B. The variable A has two levels and the variable B has three levels. There are 3 scores per "cell." A cell is a combination of one level of A and one level of B.

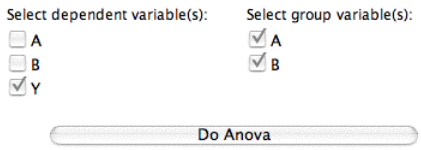
Note that Analysis Lab uses unweighted means analysis rather than the general linear model. This will give slightly different results from most other programs when there are unequal sample sizes and the degrees of freedom are greater than one.

1. Format the data so that there is one variable indicating the level of each independent variable. The variable must use consecutive integer variables. There is also a dependent variable. In the example, the variable A indicates the level of the variable A, B indicates the level of the variable B, and Y is the dependent variable.

A	B	Y
1	1	3
1	1	5
1	1	4
1	2	7
1	2	6
1	2	8
1	3	4
1	3	9
1	3	6
2	1	9
2	1	9
2	1	10
2	2	13
2	2	11
2	2	14

2	3	6
2	3	7
2	3	9

2. Paste the data into analysis lab (see the instructions for One-Way Between-Subjects).
3. Click the "Advanced" button next to the ANOVA" button.
4. Choose the dependent variable and the independent variables. The independent variables are called "group variables." Then click the "Do Anova" button.



5. The Summary table is shown below.

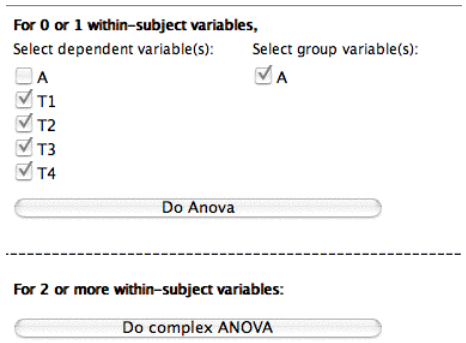
Source	df	SSQ	MS	F	p
A	1	72.0000	72.0000	32.4000	0.0001
B	2	38.1111	19.0556	8.5750	0.0049
AxB	2	20.3333	10.1667	4.5750	0.0334
Error	12	26.6667	2.2222		

One or More Between-Subject Variables and one Within-Subject Variable

1. Format the data so that a variable indicates each level of each between-subject variable. Include more than one dependent variable. For this example, there is one between-subjects variable (A) with two levels and one within-subjects variable (T) with four levels.

A	T1	T2	T3	T4
1	5	4	3	6
1	7	6	2	9
1	6	5	7	5
2	8	9	12	14
2	6	7	9	11
2	2	4	3	6

2. Paste the data into analysis lab (see the instructions for One-Way Between-Subjects).
3. Click the "Advanced" button next to the ANOVA" button.
4. Choose the dependent variables and the independent variables. The independent variables are called "group variables." In this example there is one independent variable (A) and four dependent variables ($T1$, $T2$, $T3$, and $T4$). Then click the "Do Anova" button.



5. The Summary table is shown. DV stands for dependent variable and is the effect of the within-subjects factor. $A \times DV$ is the interaction of the between-subjects variable and the within-subjects variable.

Source	df	SSQ	MS	F	p
A	1	28.1667	28.1667	1.0646	0.3605
Error	4	105.8333	26.4583		
DV	3	32.3333	10.7778	4.8199	0.0199
AxDV	3	20.8333	6.9444	3.1056	0.0670
Error	12	26.8333	2.2361		

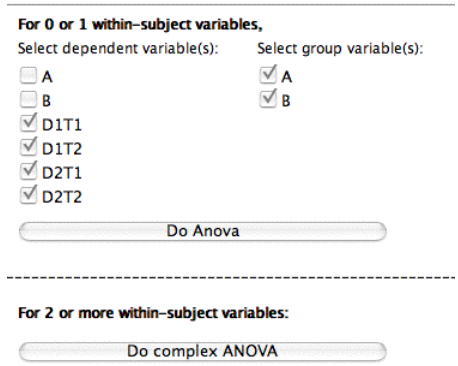
One or More Between-Subjects Variables and Two more More Within-Subjects Variables

1. Format the data so that a variable indicates each level of each between-subject variable. Include more than one dependent variable. For this example, there are two between-subjects variable (Age and Gender) and two within-subjects variable (Day and Trial). There are four groups of subjects (the combination of two levels of Age and two levels of Gender). Each subject is given two trials a day for two days. For convenience A is used as an abbreviation for age, G for Gender, D for Day, and T for Trial. The variable $D1T1$ is the score for Day 1 Trial 1, $D1T2$ for Day 1, Trial 2, etc.

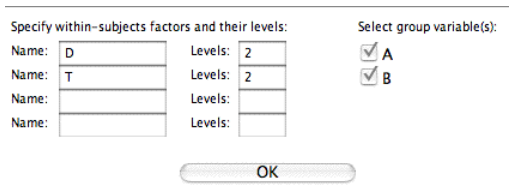
A	B	D1T1	D1T2	D2T1	D2T2
1	1	5	7	4	8
1	1	4	7	5	9
1	1	5	5	8	9
1	2	5	4	5	7
1	2	4	8	11	12
1	2	8	7	6	5
2	1	12	13	12	15
2	1	10	9	11	13
2	1	8	8	7	15
2	2	13	14	11	12
2	2	7	5	8	7
2	2	9	11	11	10

2. Paste the data into analysis lab (see the instructions for One-Way Between-Subjects).

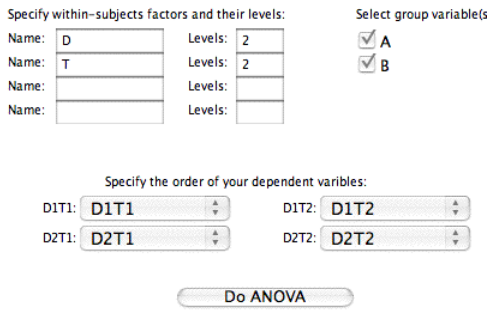
3. Click the "Advanced" button next to the ANOVA" button.
4. Choose the dependent variables and the independent variables. The independent variables are called "group variables." In this example there are two independent variable (A and B) and four dependent variables ($D1T1$, $D1T2$, $D2T1$, and $D2T2$). Then click the "Do Complex Anova" at the bottom of the window.



5. Next you name the within-subject variables and specify the number of levels of each. Since there are two days and two trials per day, both of the values are two. Analysis Lab allows up to four. In this example there are two (D and T). Also indicate that A and B are the between-subjects (group) variables.



6. Next you will be asked to associate the variable names to the levels of the variables. The way the variables were named in this example makes this a trivial exercise. However, if you used other variable names, it would take a little effort. Click the "Do ANOVA" button when finished.



7. The Summary Table is presented.

Source	df	SSQ	MS	F	p
A	1	180.1875	180.1875	12.1475	0.0083
B	1	1.6875	1.6875	0.1138	0.7446
AxB	1	9.1875	9.1875	0.6194	0.4540
Error	8	118.6667	14.8333		
D	1	22.6875	22.6875	4.2209	0.0740
AxD	1	1.0208	1.0208	0.1899	0.6745
BxD	1	3.5208	3.5208	0.6550	0.4417
AxBxD	1	3.5208	3.5208	0.6550	0.4417
Error	8	43.0000	5.3750		
T	1	20.0208	20.0208	7.6270	0.0246
AxT	1	0.5208	0.5208	0.1984	0.6678
BxT	1	11.0208	11.0208	4.1984	0.0746
AxBxT	1	0.1875	0.1875	0.0714	0.7960
Error	8	21.0000	2.6250		
DxT	1	4.6875	4.6875	3.1250	0.1151
AxDxT	1	1.0208	1.0208	0.6806	0.4333
BxDxT	1	7.5208	7.5208	5.0139	0.0555
AxBxDxT	1	2.5208	2.5208	1.6806	0.2310
Error	8	12.0000	1.5000		

15.2: Binomial Distribution

15.3: Chi Square Distribution

15.4: F Distribution

15.5: Inverse Normal Distribution

15.6: Inverse t Distribution

15.7: Normal Distribution

15.8: Power Calculator

	Power Calculator Author(s) David M. Lane Prerequisites Power Difference between means: value Standard deviation: value Sample size per group: value
Alpha level	0.05
Power (two tailed test)	
	Calculate

Fill in the fields and then press the "Caclulate" button. The power for a two-tailed t test will be displayed.

If you change a value you can press enter or the tab key to recalculate.

This calculator is based on jStat from jstat.org and is distributed under the MIT license:

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions: The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

15.9: r to Fisher z'

15.10: Studentized Range Distribution

15.11: t Distribution

Index

Glossary

Learning Objectives

- Definitions of various terms occurring throughout the textbook

A Priori Comparison / Planned Comparison

A comparison that is planned before conducting the experiment or at least before the data are examined. Also called an a priori comparison.

Absolute Deviation / Absolute Difference

The absolute value of the difference between two numbers. The absolute deviation between 5 and 3 is 2; between 3 and 5 is 2; and between -4 and 2 it is 6.

Alternative Hypothesis

In hypothesis testing, the null hypothesis and an alternative hypothesis are put forward. If the data are sufficiently strong to reject the null hypothesis, then the null hypothesis is rejected in favor of an alternative hypothesis. For instance, if the null hypothesis were that $\mu_1 = \mu_2$ then the alternative hypothesis (for a two-tailed test) would be $\mu_1 \neq \mu_2$.

Analysis of Variance

Analysis of variance is a method for testing hypotheses about means. It is the most widely-used method of statistical inference for the analysis of experimental data.

Anti-log

Taking the anti-log of a number undoes the operation of taking the log. Therefore, since $\log_{10}(1000) = 3$, the $antilog_{10}$ of 3 is $1,000$.Taking the anti-log of a number undoes the operation of taking the log. Therefore, since $\log_{10}(1000) = 3$, the $antilog_{10}$ of 3 is $1,000$. Taking the antilog of X raises the base of the logarithm in question to X.

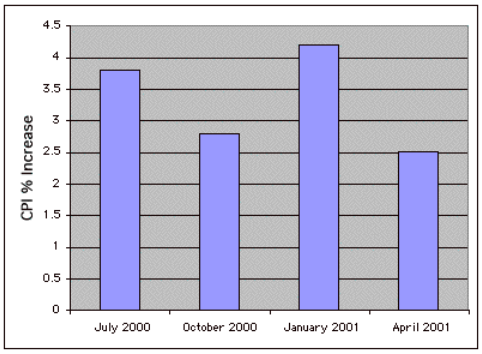
Average

1. The (arithmetic) mean
2. Any measure of central tendency

Bar Chart

A graphical method of presenting data. A bar is drawn for each level of a variable. The height of each bar contains the value of the variable. Bar charts are useful for displaying things such as frequency counts and percent increases. They are not recommended for displaying means (despite the widespread practice) since box plots present more information in the same amount of space.

An example bar chart is shown below.



Base Rate

The true proportion of a population having some condition, attribute or disease. For example, the proportion of people with schizophrenia is about 0.01. It is very important to consider the base rate when classifying people. As the saying goes, "if you hear hoofs, think horse not zebra" since you are more likely to encounter a horse than a zebra (at least in most places.)

Bayes' Theorem

Bayes' theorem considers both the prior probability of an event and the diagnostic value of a test to determine the posterior probability of the event. The theorem is shown below:

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid D')P(D')} \tag{1}$$

where $P(D|T)$ is the posterior probability of condition D given test result T, $P(T|D)$ is the conditional probability of T given D, $P(D)$ is the prior probability of D, $P(T|D')$ is the conditional probability of T given not D, and $P(D')$ is the probability of not D'.

Beta-carotene

The human body converts beta-carotene into Vitamin A, an essential component of our diets. Many vegetables, such as carrots, are good sources of beta-carotene. Studies have suggested beta-carotene supplements may provide a range of health benefits, ranging from promoting healthy eyes to preventing cancer. Other studies have found that beta-carotene supplements may increase the incidence of cancer.

Beta weight

A standardized regression coefficient.

Between-Subjects Factor / Between-Subjects Variable

Between-subject variables are independent variables or factors in which a different group of subjects is used for each level of the variable. If an experiment is conducted comparing four methods of teaching vocabulary and if a different group of subjects is used for each of the four teaching methods, then teaching method is a between-subjects variable.

Between-Subjects Factor / Between-Subjects Variable

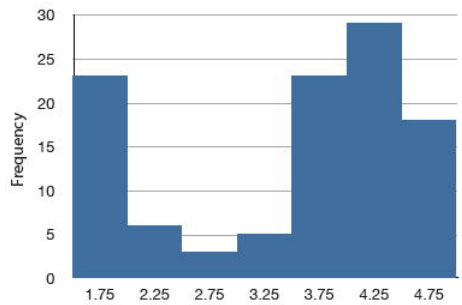
Between-subject variables are independent variables or factors in which a different group of subjects is used for each level of the variable. If an experiment is conducted comparing four methods of teaching vocabulary and if a different group of subjects is used for each of the four teaching methods, then teaching method is a between-subjects variable.

Bias

1. A sampling method is biased if each element does not have an equal chance of being selected. A sample of internet users found reading an online statistics book would be a biased sample of all internet users. A random sample is unbiased. Note that possible bias refers to the sampling method, not the result. An unbiased method could, by chance, lead to a very non-representative sample.
2. An estimator is biased if it systematically overestimates or underestimates the parameter it is estimating. In other words, it is biased if the mean of the sampling distribution of the statistic is not the parameter it is estimating, The sample mean is an unbiased estimate of the population mean. The mean squared deviation of sample scores from their mean is a biased estimate of the variance since it tends to underestimate the population variance.

Bimodal Distribution

A distribution with two distinct peaks. An example is shown below.



Binomial Distribution

A probability distribution for independent events for which there are only two possible outcomes such as a coin flip. If one of the two outcomes is defined as a success, then the probability of exactly x successes out of N trials (events) is given by:

$$P(x) = \frac{N!}{x!(N-x)!}\pi^x(1-\pi)^{N-x} \tag{2}$$

Bivariate

Bivariate data is data for which there are two variables for each observation. As an example, the following bivariate data show the ages of husbands and wives of 10 married couples.

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

Bonferroni correction

In general, to keep the familywise error rate (FER) at or below 0.05, the per-comparison error rate ($PCER$) should be:

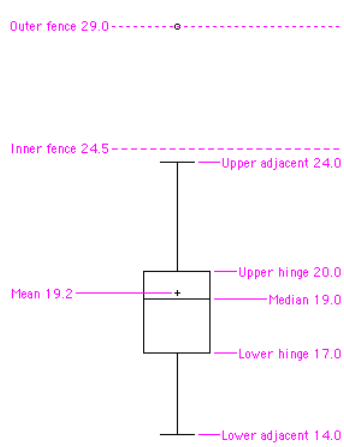
$$PCER = 0.05/c \tag{3}$$

where c is the number of comparisons. More generally, to insure that the FER is less than or equal to alpha, use

$$PCER = alpha/c \tag{4}$$

Box Plot

One of the more effective graphical summaries of a data set, the box plot generally shows mean, median, 25th and 75th percentiles, and outliers. A standard box plot is composed of the median, upper hinge, lower hinge, upper adjacent value, lower adjacent value, outside values, and far out values. An example is shown below. Parallel box plots are very useful for comparing distributions. See also: step, H-spread.



Center (of a distribution) / Central Tendency

The center or middle of a distribution. There are many measures of central tendency. The most common are the mean, median, and, mode. Others include the trimean, trimmed mean, and geometric mean.)

Class Frequency

One of the components of a histogram, the class frequency is the number of observations in each class interval.

Class Interval / Bin Width

Also known as bin width, the class interval is a division of data for use in a histogram. For instance, it is possible to partition scores on a 100 point test into class intervals of $1-25$, $26-49$, $50-74$ and $75-100$.

Comparison Among Means/ Contrast Among Means

A method for testing differences among means for significance. For example, one might test whether the difference between $\text{Mean } 1$ and the average of $\text{Mean } 2$ and $\text{Mean } 3$ is significantly different.

Conditional Probability

The probability that event A occurs given that event B has already occurred is called the conditional probability of A given B. Symbolically, this is written as $P(A|B)$. The probability it rains on Monday given that it rained on Sunday would be written as $\text{P}(\text{Rain on Monday} \mid \text{Rain on Sunday})$.

Confidence Interval

A confidence interval is a range of scores likely to contain the parameter being estimated. Intervals can be constructed to be more or less likely to contain the parameter: 95% of 95% confidence intervals contain the estimated parameter whereas 99% of 99% confidence intervals contain the estimated parameter. The wider the confidence interval, the more uncertainty there is about the value of the parameter.

Confounding

Two or more variables are confounded if their effects cannot be separated because they vary together. For example, if a study on the effect of light inadvertently manipulated heat along with light, then light and heat would be confounded.

Cook's D

Cook's D is a measure of the influence of an observation in regression and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question.

Constant

A value that does not change. Values such as pi, or the mass of the Earth are constants. Compare with variables.

Continuous Variables

Variables that can take on any value in a certain range. Time and distance are continuous; gender, SAT score and “time rounded to the nearest second” are not. Variables that are not continuous are known as discrete variables. No measured variable is truly continuous; however, discrete variables measured with enough precision can often be considered continuous for practical purposes.

Correlated Pairs t test / Related Pairs t test

A test of the difference between means of two conditions when there are pairs of scores. Typically, each pair of scores is from a different subject.

Counterbalanced

Counterbalancing is a method of avoiding confounding among variables. Consider an experiment in which subjects are tested on both an auditory reaction time task (in which subjects respond to an auditory stimulus) and a visual reaction time task (in which

subjects respond to a visual stimulus). Half of the subjects are given the visual task first and the other half of the subjects are given the auditory task first. That way, there is no confounding of order of presentation and task.

Consumer Price Index - CPI

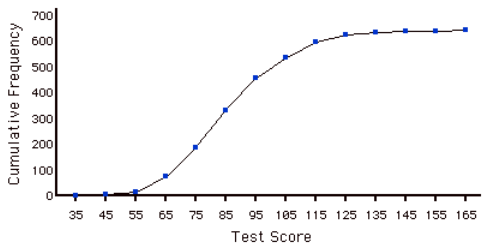
Also known as the cost of living index, the CPI is a financial statistic that measures the change in price of a representative group of goods over time.

Criterion Variable

In regression analysis (such as linear regression) the criterion variable is the variable being predicted. In general, the criterion variable is the dependent variable.

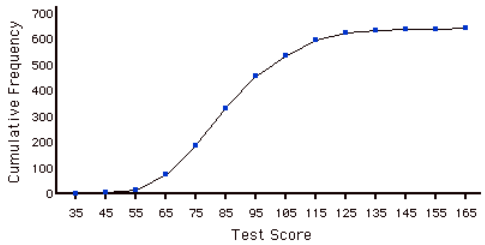
Cumulative Frequency Distribution

A distribution showing the number of observations less than or equal to values on the X-axis. The following graph shows a cumulative distribution for scores on a test.



Cumulative Frequency Polygon

A frequency polygon whose vertices represent the sum of all previous class frequencies of the data. For example, if a frequency polygon had vertices of $5, 8, 3, 7, 10$, the cumulative frequency polygon on the same data would have vertices of $5, 13, 16, 23, 33$ As another example, a cumulative frequency distribution for scores on a psychology test is shown below.



Data

A collection of values to be used for statistical analysis.

Dependent Variable

A variable that measures the experimental outcome. In most experiments, the effects of the independent variable on the dependent variables are observed. For example, if a study investigated the effectiveness of an experimental treatment for depression, then the measure of depression would be the dependent variable.

Synonym: dependent measure

Descriptive Statistics

1. The branch of statistics concerned with describing and summarizing data.
2. A set of statistics such as the mean, standard deviation, and skew that describe a distribution.

Deviation Scores

Scores that are expressed as differences (deviations) from some value, usually the mean. To convert data to deviation scores typically means to subtract the mean score from each other score. Thus, the values 1, 2, and 3 in deviation-score form would be computed by subtracting the mean of 2 from each value and would be $-1, 0, 1$.

Degrees of Freedom - df

The degrees of freedom of an estimate is the number of independent pieces of information that go into the estimate. In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. For example, to estimate the population variance, one must first estimate the population mean. Therefore, if the estimate of variance is based on N observations, there are $N-1$ degrees of freedom.

Discrete Variable

Variables that can only take on a finite number of values are called "discrete variables." All qualitative variables are discrete. Some quantitative variables are discrete, such as performance rated as 1,2,3,4, or 5, or temperature rounded to the nearest degree. Sometimes, a variable that takes on enough discrete values can be considered to be continuous for practical purposes. One example is time to the nearest millisecond.

Variables that can take on an infinite number of possible values are called "continuous variables."

Distribution / Frequency Distribution

The distribution of empirical data is called a frequency distribution and consists of a count of the number of occurrences of each value. If the data are continuous, then a grouped frequency distribution is used. Typically, a distribution is portrayed using a frequency polygon or a histogram.

Mathematical equations are often used to define distributions. The normal distribution is, perhaps, the best known example. Many empirical distributions are approximated well by mathematical distributions such as the normal distribution.

Expected Value

The expected value of a statistic is the mean of the sampling distribution of the statistic. It can be loosely thought of as the long-run average value of the statistic.

Factor / Independent Variables

Variables that are manipulated by the experimenter, as opposed to dependent variables. Most experiments consist of observing the effect of the independent variable(s) on the dependent variable(s).

Factorial Design

In a factorial design, each level of each independent variable is paired with each level of each other independent variable. Thus, a 2×3 factorial design consists of the 6 possible combinations of the levels of the independent variables.

False Positive

A false positive occurs when a diagnostic procedure returns a positive result while the true state of the subject is negative. For example, if a test for strep says the patient has strep when in fact he or she does not, then the error in diagnosis would be called a false positive. In some contexts, a false positive is called a false alarm. The concept is similar to a Type I error in significance testing.

Familywise Error Rate

When a series of significance tests is conducted, the familywise error rate (FER) is the probability that one or more of the significance tests results in a Type I error.

Far Out Value

One of the components of a box plot, far out values are those that are more than 2 steps from the nearest hinge. They are beyond the outer fences.

Favorable Outcome

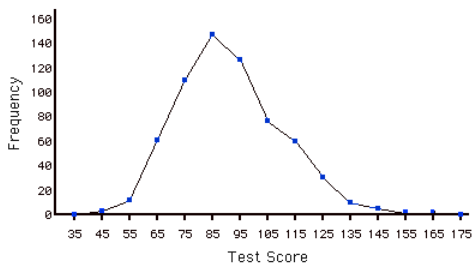
A favorable outcome is the outcome of interest. For example one could define a favorable outcome in the flip of a coin as a head. The term "favorable outcome" does not necessarily mean that the outcome is desirable – in some experiments, the favorable outcome could be the failure of a test, or the occurrence of an undesirable event.

Frequency Distribution

For a discrete variable, a frequency distribution consists of the distribution of the number of occurrences for each value of the variable. For a continuous variable, it is the number of occurrences for a variety of ranges of variables.

Frequency Polygon

A frequency polygon is a graphical representation of a distribution. It partitions the variable on the x-axis into various contiguous class intervals of (usually) equal widths. The heights of the polygon's points represent the class frequencies.



Frequency Table

A table containing the number of occurrences in each class of data; for example, the number of each color of M&Ms in a bag. Frequency tables often used to create histograms and frequency polygons. When a frequency table is created for a quantitative variable, a grouped frequency table is generally used.

Grouped Frequency Table

A grouped frequency table shows the number of values for various ranges of scores. Below is shown a grouped frequency table for response times (in milliseconds) for a simple motor task.

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Geometric Mean

The geometric mean is a measure of central tendency. The geometric mean of n numbers is obtained by multiplying all of them together, and then taking the nth root of them. For example, for the numbers 1, 10, and 100, the product of all the numbers is: $1 \times 10 \times 100 = 1,000$. Since there are three numbers, we take the cubed root of the product $(1,000)$ which is equal to 10.

Graphs / Graphics

Graphs are often the most effective way to describe distributions and relationships among variables. Among the most common graphs are histograms, box plots, bar charts, and scatterplots.

Grouped Frequency Distribution

A grouped frequency distribution is a frequency distribution in which frequencies are displayed for ranges of data rather than for individual values. For example, the distribution of heights might be calculated by defining one-inch ranges. The frequency of individuals with various heights rounded off to the nearest inch would then be tabulated.

Harmonic Mean

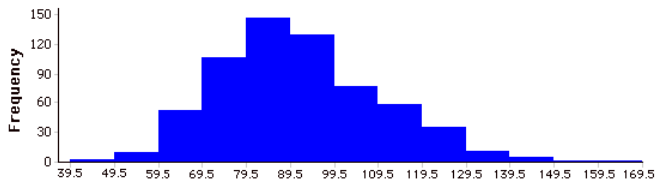
The harmonic mean of n numbers (X_1 to X_n) is computed using the following formula:

$$Harmonic\ Mean = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \tag{5}$$

Often the harmonic mean of sample sizes is computed.

Histogram

A histogram is a graphical representation of a distribution . It partitions the variable on the x-axis into various contiguous class intervals of (usually) equal widths. The heights of the bars represent the class frequencies.



History Effect

A problem of confounding where the passage of time, and not the variable of interest, is responsible for observed effects.

Homogeneity of variance

The assumption that the variances of all the populations are equal.

Homoscedasticity

In linear regression, the assumption that the variance around the regression line is the same for all values of the predictor variable.

H-spread

One of the components of a box plot, the H-spread is the difference between the upper hinge and the lower hinge.

iMac

A line of computers released by Apple in 1998, that tried to make computers both more accessible (greater ease of use) and fashionable (they came in a line of designer colors.)

Independence

Two variables are said to be independent if the value of one variable provides no information about the value of the other variable. These two variables would be uncorrelated so that Pearson's r would be 0.

Two events are independent if the probability the second event occurring is the same regardless of whether or not the first event occurred.

Independent Events

Events A and B are independent events if the probability of Event B occurring is the same whether or not Event A occurs. For example, if you throw two dice, the probability that the second die comes up 1 is independent of whether the first die came up 1. Formally, this can be stated in terms of conditional probabilities:

$$P(A|B) = P(A) \quad (6)$$
$$P(B|A) = P(B)$$

Inference

The act of drawing conclusions about a population from a sample.

Inferential Statistics

The branch of statistics concerned with drawing conclusions about a population from a sample. This is generally done through random sampling, followed by inferences made about central tendency, or any of a number of other aspects of a distribution.

Influence

Influence refers to the degree to which a single observation in regression influences the estimation of the regression parameters. It is often measured in terms how much the predicted scores for other observations would differ if the observation in question were not included.

Inner Fence

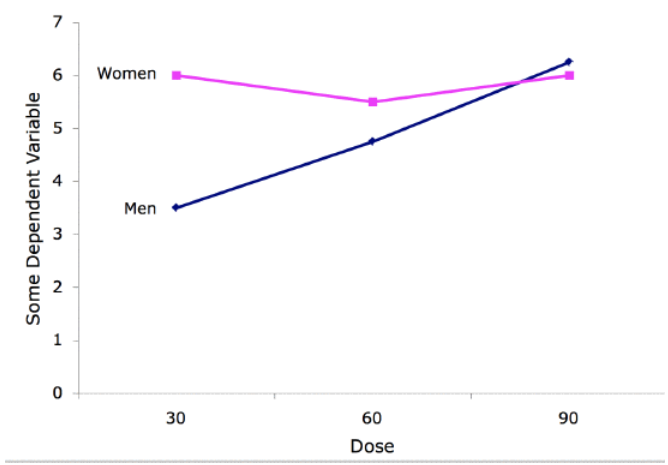
In a box plot, the lower inner fence is one step below the lower hinge while the upper inner fence is one step above the upper hinge.

Interaction

Two independent variables interact if the effect of one of the variables differs depending on the level of the other variable.

Interaction Plot

An interaction plot displays the levels of one variable on the X axis and has a separate line for the means of each level of the other variable. The Y axis is the dependent variable. A look at this graph shows that the effect of dosage is different for males than it is for females.



Interquartile Range

The Interquartile Range (IQR) is the (75th percentile – 25th percentile). It is a robust measure of variability.

Interval Estimate

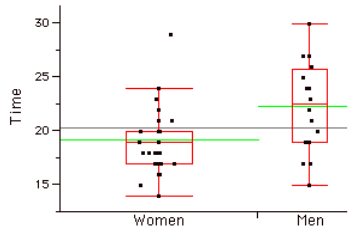
An interval estimate is a range of scores likely to contain the estimated parameter.

Interval Scales

One of 4 Levels of Measurement, interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10 degree interval has the same physical meaning (in terms of the kinetic energy. Unlike ratio scales, interval scales do not have a true zero point.

Jitter

When points in a graph are jittered, the are moved horizontally so that all the points can be seen and none are hidden due to overlapping values. An example is shown below:



Kurtosis

Kurtosis measures how fat or thin the tails of a distribution are relative to a normal distribution. It is commonly defined as:

$$\sum \frac{(X-\mu)^4}{\sigma^4} - 3 \tag{7}$$

Distributions with long tails are called leptokurtic; distributions with short tails are called platykurtic. Normal distributions have zero kurtosis.

Leptokurtic

A distribution with long tails relative to a normal distribution is leptokurtic.

Level / Level of a variable / Level of a factor

When a factor consists of various treatment conditions, each treatment condition is considered a level of that factor. For example, if the factor were drug dosage, and three doses were tested, then each dosage would be one level of the factor and the factor would have three levels.

Levels of Measurement

Measurement scales differ in their level of measurement. There are four common levels of measurement:

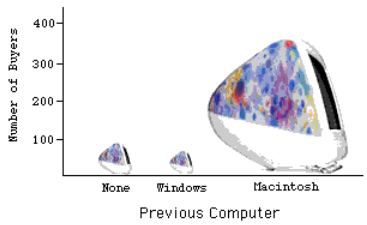
1. Nominal scales are only labels.
2. Ordinal Scales are ordered but are not truly quantitative. Equal intervals on the ordinal scale do not imply equal intervals on the underlying trait.
3. Interval scales are are ordered and equal intervals equal intervals on the underlying trait. However, interval scales do not have a true zero point.
4. Ratio scales are interval scales that do have a true zero point. With ratio scales, it is sensible to talk about one value being twice as large as another, for example.

Leverage

Leverage is a factor affecting the influence of an observation in regression. Leverage is based on how much the observation's value on the predictor variable differs from the mean of the predictor variable. The greater an observation's leverage, the more potential it has to be an influential observation.

Lie Factor

Many problems can arise when fancy graphs are used over plain ones. Distortions can occur when the heights of objects are used to indicate the value because most people will pay attention to the areas of the objects rather than their height. The lie factor is the ratio of the effect apparent in the graph to actual effect in the data; if it deviates by more than 0.05 from 1, the graph is generally unacceptable. The lie factor in the following graph is almost 6.



Lies

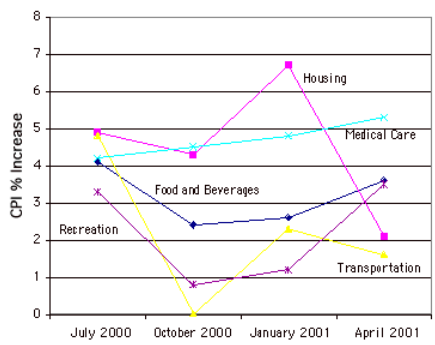
There are three types of lies:

i. regular lies
ii. damned lies
iii. statistics

This is according to Benjamin Disraeli as quoted by Mark Twain.

Line Graph

Essentially a bar graph in which the height of each par is represented by a single point, with each of these points connected by a line. Line graphs are best used to show change over time, and should never be used if your X-axis is not an ordered variable. An example is shown below.



Linear Combination

A linear combination of variables is a way of creating a new variable by combining other variables. A linear combination is one in which each variable is multiplied by a coefficient and the are products summed. For example, if

$$Y = 3X_1 + 2X_2 + 0.5X_3 \tag{8}$$

then Y is a linear combination of the variables X_1, X_2, and X_3.

Linear Regression

Linear regression is a method for predicting a criterion variable from one or more predictor variable. In simple regression, the criterion is predicted from a single predictor variable and the best-fitting straight line is of the form:

$$Y' = bX + A \tag{9}$$

where Y' is the predicted score, X is the predictor variable, b is the slope, and A is the Y intercept. Typically, the criterion for the "best fitting" line is the line for which the sum of the squared errors of prediction is minimized. In multiple regression, the criterion is predicted from two or more predictor variables.

Linear Relationship

There is a perfect linear relationship between two variables if a scatterplot of the points falls on a straight line. The relationship is linear even if the points diverge from the line as long as the divergence is random rather than being systematic.

Linear Transformation

A linear transformation is any transformation of a variable that can be achieved by multiplying it by a constant, and then adding a second constant. If Y is the transformed value of X, then $Y = aX + b$. The transformation from degrees Fahrenheit to degrees Centigrade is linear and is done using the formula:

$$C = 0.55556F - 17.7778 \tag{10}$$

Logarithm - Log

The logarithm of a number is the power the base of the logarithm has to be raised to in order to equal the number. If the base of the logarithm is 10 and the number is $1,000$ then the log is 3 since 10 has to be raised to the 3rd power to equal $1,000$.

Lower Adjacent Value

A component of a box plot, the lower adjacent value is smallest value in the data above the inner lower fence.

Lower Hinge

A component of a box plot, the lower hinge is the 25th percentile. The upper hinge is the 75th percentile.

M & M's

A type of candy consisting of chocolate inside a shell. M & M's come in a variety of colors.

Main Effect

A main effect of an independent variable is the effect of the variable averaging over all levels of the other variable(s). For example, in a design with age and gender as factors, the main effect of gender would be the difference between the genders averaging across all ages used in the experiment.

Margin of Error

When a statistic is used to estimate a parameter, it is common to compute a confidence interval. The margin of error is the difference between the statistic and the endpoints of the interval. For example, if the statistic were 0.6 and the confidence interval ranged from 0.4 to 0.8, then the margin of error would be 0.20. Unless otherwise specified, the 95% confidence interval is used.

Marginal Mean

In a design with two factors, the marginal means for one factor are the means for that factor averaged across all levels of the other factor. In the table shown below, the two factors are "Relationship" and "Companion Weight." The marginal means for each of the two levels of Relationship (Girl Friend and Acquaintance) are computed by averaging across the two levels of Companion Weight. Thus, the marginal mean for Acquaintance of 6.37 is the mean of 6.15 and 6.59.

		Companion Weight		
		Obese	**Typical**	**Marginal Mean**
Relationship	**Girl Friend**	5.65	6.19	5.92
	Acquaintance	6.15	6.59	6.37
	Marginal Mean	5.90	6.39	

Mean / Arithmetic Mean

Also known as the arithmetic mean, the mean is typically what is meant by the word “average.” The mean is perhaps the most common measure of central tendency. The mean of a variable is given by **(the sum of all its values)/(the number of values)**. For example, the mean of 4, 8, and 9 is 7. The sample mean is written as M, and the population mean as the Greek letter mu (μ). Despite its popularity, the mean may not be an appropriate measure of central tendency for skewed distributions, or in situations with outliers.

Median

The median is a popular measure of central tendency. It is the 50th percentile of a distribution. To find the median of a number of values, first order them, then find the observation in the middle: the median of $5, 2, 7, 9, 4$ is 5. (Note that if there is an even number of values, one takes the average of the middle two: the median of $4, 6, 8, 10$ is 7.) The median is often more appropriate than the mean in skewed distributions and in situations with outliers.

Miss

Misses occur when a diagnostic test returns a negative result, but the true state of the subject is positive. For example, if a person has strep throat and the diagnostic test fails to indicate it, then a miss has occurred. The concept is similar to a Type II error in significance testing.

Mode

The mode is a measure of central tendency. It is the most frequent value in a distribution: the mode of $3, 4, 4, 5, 5, 5, 8$ is 5. Note that the mode may be very different from the mean and the median.

Multiple Regression

Multiple regression is linear regression in which two or more predictor variables are used to predict the criterion.

Negative Association

There is a negative association between variables X and Y if smaller values of X are associated with larger values of Y and larger values of X are associated with smaller values of Y.

Nominal Scale

A nominal scale is one of four commonly-used Levels of Measurement. No ordering is implied, and addition/subtraction and multiplication/division would be inappropriate for a variable on a nominal scale. {Female, Male} and {Buddhist, Christian, Hindu, Muslim} have no natural ordering (except alphabetic). Occasionally, numeric values are nominal: for instance, if a variable were coded as $\text{Female} = 1$, $\text{Male} = 2$, the set $1, 2$ is still nominal.

Non representative

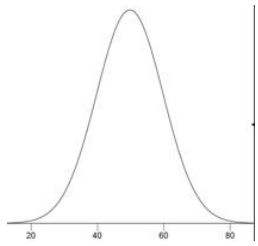
A non-representative sample is a sample that does not accurately reflect the population.

Normal Distribution

One of the most common continuous distributions, a normal distribution is sometimes referred to as a "bell-shaped distribution." If μ is the distribution mean, and σ the standard deviation, then the height (ordinate) normal distribution is given by

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{11}$$

A graph of a normal distribution with a mean of 50 and a standard deviation of 10 is shown below.



If the mean is 0 and the standard deviation is 1, the distribution is referred to as the "standard normal distribution."

Null Hypothesis

A null hypothesis is a hypothesis tested in significance testing. It is typically the hypothesis that a parameter is zero or that a difference between parameters is zero. For example, the null hypothesis might be that the difference between population means is zero. Experimenters typically design experiments to allow the null hypothesis to be rejected.

Omninus Null Hypothesis

The null hypothesis that all population means are equal.

One-Tailed Test / One-Tailed Probability / Directional Test

The last step in significance testing involves calculating the probability that a statistic would differ as much or more from the parameter specified in the null hypothesis as does the statistics obtained in the experiment.

A probability computed considering differences in only one direction, such as the statistic is larger than the parameter, is called a one-tailed probability. For example, if a parameter is 0 and the statistic is 12, a one-tailed probability (the positive tail) would be the probability of a statistic being ≥ 12. Compare with the two-tailed probability which would be the probability of being either ≤ -12 or ≥ 12.

Ordinal Scale

One of four commonly used levels of measurement, an ordinal scale is a set of ordered values. However, there is no set distance between scale values. For instance, for the scale: (Very Poor, Poor, Average, Good, Very Good) is an ordinal scale. You can assign numerical values to an ordinal scale: rating performance such as 1 for "Very Poor," 2 for "Poor," etc, but there is no assurance that the difference between a score of 1 and 2 means the same thing as the difference between a score of 2 and 3.

Orthogonal Comparisons

When comparisons among means provide completely independent information, the comparisons are called "orthogonal." If an experiment with four groups were conducted, then a comparison of Groups 1 and 2 would be orthogonal to a comparison of Groups 3 and 4 since there is nothing in the comparison of Groups 1 and 2 that provides information about the comparison of Groups 3 and 4.

Outer Fence

In a box plot, the lower outer fence is 2 steps below the lower hinge whereas the upper inner fence is 2 steps above the upper hinge.

Outlier

Outliers are atypical, infrequent observations; values that have an extreme deviation from the center of the distribution. There is no universally-agreed on criterion for defining an outlier, and outliers should only be discarded with extreme caution. However, one should always assess the effects of outliers on the statistical conclusions.

Outside Value

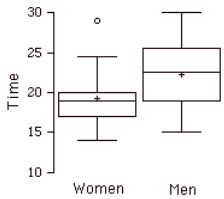
A component of a box plot, outside values are more than 1 step beyond the nearest hinge. They are beyond an inner fence but not beyond an outer fence.

Pairwise Comparisons

Pairwise comparisons are comparisons between pairs of means.

Parallel Box Plots

Two or more box plots drawn on the same Y-axis. These are often useful in comparing features of distributions. An example portraying the times it took samples of women and men to do a task is shown below.



Parameter

A value calculated in a population. For example, the mean of the numbers in a population is a parameter. Compare with a statistic, which is a value computed in a sample to estimate a parameter.

Partial Slope

The partial slope in multiple regression is the slope of the relationship between a predictor variable that is independent of the other predictor variables and the criterion. It is also the regression coefficient for the predictor variable in question.

Pearson's r / Pearson's Product Moment Correlation / Pearson's Correlation

Pearson's correlation is a measure of the strength of the linear relationship between two variables. It ranges from -1 for a perfect negative relationship to $+1$ for a perfect positive relationship. A correlation of 0 means that there is no linear relationship.

Percentile

There is no universally accepted definition of a percentile. Using the 65th percentile as an example, some statisticians define the 65th percentile as the lowest score that is greater than 65% of the scores. Others have defined the 65th percentile as the lowest score that is greater than or equal to 65% of the scores. A more sophisticated definition is given below. The first step is to compute the rank (R) of the percentile in question. This is done using the following formula:

$$R = \frac{P}{100} \times (N+1) \tag{12}$$

where P is the desired percentile and N is the number of numbers. If R is an integer, then the Pth percentile is the number with rank R. When R is not an integer, we compute the Pth percentile by interpolation as follows:

1. Define IR as the integer portion of R (the number to the left of the decimal point).
2. Define FR as the fractional portion or R.
3. Find the scores with Rank IR and with Rank $IR+1$.
4. Interpolate by multiplying the difference between the scores by FR and add the result to the lower score.

Per-Comparison Error Rate

The per-comparison error rate refers to the **Type I** error rate of any one significance test conducted as part of a series of significance tests. Thus, if 10 significance tests were each conducted at 0.05 significance level, then the per-comparison error rate would be 0.05. Compare with the familywise error rate.

Pie Chart

A graphical representation of data, the pie chart shows relative frequencies of classes of data. It is a circle cut into a number of wedges, one for each class, with the area of each wedge proportional to its relative frequency. Pie charts are only effective for a small number of classes, and are one of the less effective graphical representations.

Placebo

A device used in clinical trials, the placebo is visually indistinguishable from the study medication, but in reality has no medical effect (often, a sugar pill). A group of subjects chosen randomly takes the placebo, the others take one or another type of medication. This is done to prevent confounding the medical and psychological effects of the drug. Even a sugar pill can lead some patients to report improvement and side effects.

Planned Comparison / A Priori Comparison

A comparison that is planned before conducting the experiment or at least before the data are examined. Also called an a priori comparison.

Platykurtic

A distribution with short tails relative to a normal distribution is platykurtic.

Point Estimate

When a parameter is being estimated, the estimate can be either a single number or it can be a range of numbers such as in a confidence interval. When the estimate is a single number, the estimate is called a "point estimate."

Polynomial Regression

Polynomial regression is a form of multiple regression in which powers of a predictor variable instead of other predictor variables are used. In the following example, the criterion (Y) is predicted by X, X^2 and, X^3.

$$Y = b_1X + b_2X^2 + b_3X^3 + A \tag{13}$$

Population

A population is the complete set of observations a researcher is interested in. Contrast this with a sample which is a subset of a population. A population can be defined in a manner convenient for a researcher. For example, one could define a population as all girls in fourth grade in Houston, Texas. Or, a different population is the set of all girls in fourth grade in the United States. Inferential statistics are computed from sample data in order to make inferences about the population.

Positive Association

There is a positive association between variables X and Y if smaller values of X are associated with smaller values of Y and larger values of X are associated with larger values of Y.

Posterior Probability

The posterior probability of an event is the probability of the event computed following the collection of new data. One begins with a prior probability of an event and revises it in the light of new data. For example, if 0.01 of a population has schizophrenia then the probability that a person drawn at random would have schizophrenia is 0.01. This is the prior probability. If you then learn that that their score on a personality test suggests the person is schizophrenic, you would adjust your probability accordingly. The adjusted probability is the posterior probability.

Power

In significance testing, power is the probability of rejecting a false null hypothesis.

Precision

A statistic's precision concerns to how close it is expected to be to the parameter it is estimating. Precise statistics are vary less from sample to sample. The precision of a statistic is usually defined in terms of it standard error.

Predictor Variable

A predictor variable is a variable used in regression to predict another variable. It is sometimes referred to as an independent variable if it is manipulated rather than just measured.

Prior Probability

The prior probability of an event is the probability of the event computed before the collection of new data. One begins with a prior probability of an event and revises it in the light of new data. For example, if 0.01 of a population has schizophrenia then the probability that a person drawn at random would have schizophrenia is 0.01. This is the prior probability. If you then learn that that there score on a personality test suggests the person is schizophrenic, you would adjust your probability accordingly. The adjusted probability is the posterior probability.

Probability Density Function

For a discrete random variable, a probability distribution contains the probability of each possible outcome. However, for a continuous random variable, the probability of any one outcome is zero (if you specify it to enough decimal places). A probability density function is a formula that can be used to compute probabilities of a range of outcomes for a continuous random variable. The sum of all densities is always 1.0 and the value of the function is always greater or equal to zero.

Probability Distribution

For a discrete random variable, a probability distribution contains the probability of each possible outcome. The sum of all probabilities is always 1.0.

Probability Value / p value

In significance testing, the probability value (sometimes called the p value) is the probability of obtaining a statistic as different or more different from the parameter specified in the null hypothesis as the statistic obtained in the experiment. The probability value is computed assuming the null hypothesis is true. The lower the probability value, the stronger the evidence that the null hypothesis is false. Traditionally, the null hypothesis is rejected if the probability value is below 0.05.

Probability values can be either one tailed or two tailed.

Qualitative Variables / Categorical Variable

Also known as categorical variables, qualitative variables are variables with no natural sense of ordering. They are therefore measured on a nominal scale. For instance, hair color (Black, Brown, Gray, Red, Yellow) is a qualitative variable, as is name (Adam, Becky, Christina, Dave . . .). Qualitative variables can be coded to appear numeric but their numbers are meaningless, as in $\text{male}=1$, $\text{female}=2$. Variables that are not qualitative are known as quantitative variables.

Quantitative Variable

Variables that are measured on a numeric or quantitative scale. Ordinal, interval and ratio scales are quantitative. A country's population, a person's shoe size, or a car's speed are all quantitative variables. Variables that are not quantitative are known as qualitative variables.

Random Assignment

Random assignment occurs when the subjects in an experiment are randomly assigned to conditions. Random assignment prevents systematic confounding of treatment effects with other variables.

Random Sampling / Simple Random Sampling

The process of selecting a subset of a population for the purposes of statistical inference. Random sampling means that every member of the population is equally likely to be chosen.

Range

The difference between the maximum and minimum values of a variable or distribution. The range is the simplest measure of variability.

Ratio Scale

One of the four basic levels of measurement, a ratio scale is a numerical scale with a true zero point and in which a given size interval has the same interpretation for the entire scale. Weight is a ratio scale, Therefore, it is meaningful to say that a 200 pound person weighs twice as much as a 100 pound person.

Regression

Regression means "prediction." The regression of Y on X means the prediction of Y by X.

Regression Coefficient

A regression coefficient is the slope of the regression line in simple regression or the partial slope in multiple regression.

Regression Line

In linear regression, the line of best fit is called the regression line.

Relative frequency

The proportion of observations falling into a given class. For example, if a bag of 55 M & M's has 11 green M&M's, then the frequency of green M&M's is 11 and the relative frequency is $11/55 = 0.20$ Relative frequencies are often used in histograms, pie charts, and bar graphs.

Relative Frequency Distribution

A relative frequency distribution is just like a frequency distribution except that it consists of the proportions of occurrences instead of the numbers of occurrences for each value (or range of values) of a variable.

Reliability

Although there are many ways to conceive of the reliability of a test, the classical way is to define the reliability as the correlation between two parallel forms of the test. When defined this way, the reliability is the ratio of true score variance to test score variance. Chronbach's α is a common measure of reliability.

Repeated-Measures Factor / Repeated-Measures Variable / Within-Subjects Factor /Within-Subjects Variable

A within-subjects variable is an independent variable that is manipulated by testing each subject at each level of the variable. Compare with a between-subjects variable in which different groups of subjects are used for each level of the variable.

Representative Sample

A representative sample is a sample chosen to match the qualities of the population from which it is drawn. With a large sample size, random sampling will approximate a representative sample; stratified random sampling can be used to make a small sample more representative.

Robust

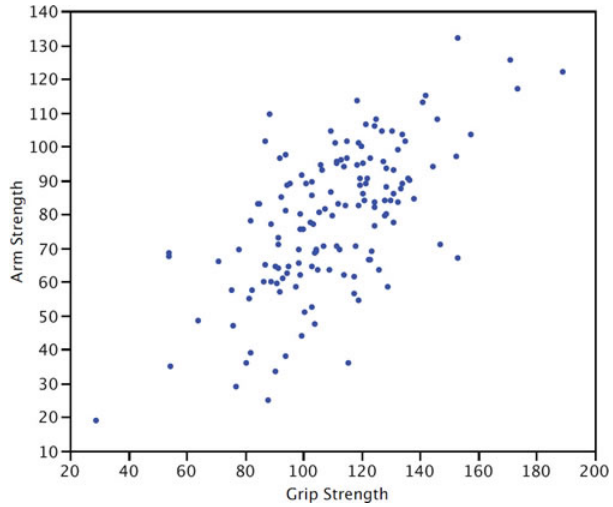
Something is robust if it holds up well in the face of adversity. A measure of central tendency or variability is considered robust if it is not greatly affected by a few extreme scores. A statistical test is considered robust if it works well in spite of moderate violations of the assumptions on which it is based.

Sample

A sample is a subset of a population, often taken for the purpose of statistical inference. Generally, one uses a random sample.

Scatter Plot

A scatter plot of two variables shows the values of one variable on the Y axis and the values of the other variable on the X axis. Scatter plots are well suited for revealing the relationship between two variables. The scatter plot shown below illustrates the relationship between grip strength and arm strength in a sample of workers.



Semi-Interquartile Range

The semi interquartile range is the interquartile range divided by 2. It is a robust measure of variability. The Interquartile Range is the (75th percentile – 25th percentile).

Significance Level / A level

In significance testing, the significance level is the highest value of a probability value for which the null hypothesis is rejected. Common significance levels are 0.05 and 0.01. If the 0.05 level is used, then the null hypothesis is rejected if the probability value is less than or equal to 0.05.

Significance Testing / Hypothesis Testing / Significant Difference

A statistical procedure that tests the viability of the null hypothesis. If data (or more extreme data) are very unlikely given that the null hypothesis is true, then the null hypothesis is rejected. If the data or more extreme data are not unlikely, then the null hypothesis is not rejected. If the null hypothesis is rejected, then the result of the test is said to be significant. A statistically significant effect does not mean the effect is important.

Simple Effect

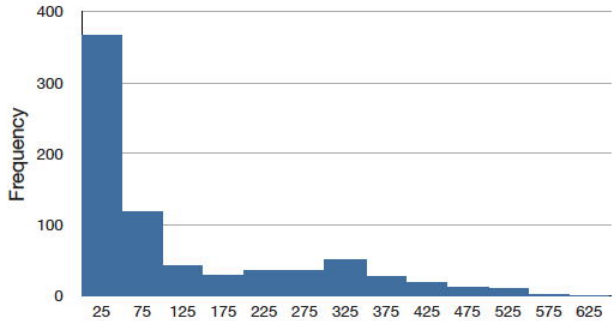
The simple effect of a factor is the effect of that factor at a single level of another factor. For example, in a design with age and gender as factors, the effect of age for females would be one of the simple effects of age.

Simple Regression

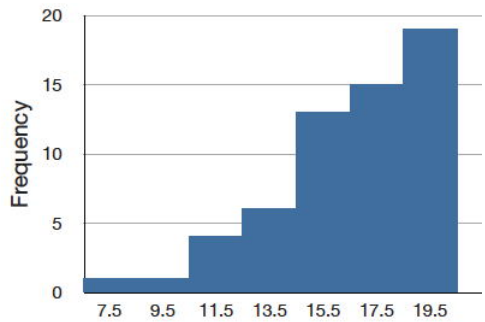
Simple regression is linear regression in which one more predictor variable is used to predict the criterion.

Skew

A distribution is skewed if one tail extends out further than the other. A distribution has positive skew (is skewed to the right) if the tail to the right is longer. See the graph below for an example.

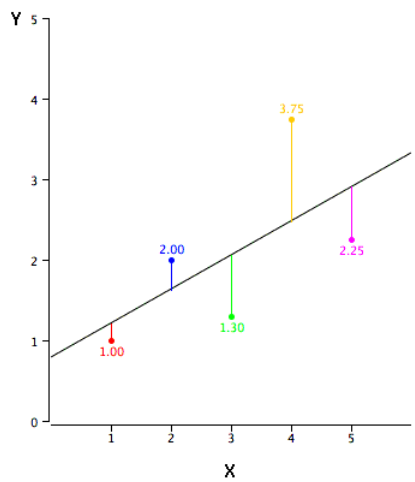


A distribution has a negative skew (is skewed to the left) if the tail to the left is longer. See the graph below for an example.



Slope

The slope of a line is the change in Y for each change of one unit of X. It is sometimes defined as "rise over run" which is the same thing. The slope of the black line in the graph is 0.425 because the line increases by 0.425 each time X increases by 1.0.



Squared Deviation

A squared deviation is the difference between two values, squared. The number that minimizes the sum of squared deviations for a variable is its mean.

Standard Deviation

The standard deviation is a widely used measure of variability. It is computed by taking the square root of the variance. An important attribute of the standard deviation as a measure of variability is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score.

Standard Error

The standard error of a statistic is the standard deviation of the sampling distribution of that statistic. For example, the standard error of the mean is the standard deviation of the sampling distribution of the mean. Standard errors play a critical role in constructing confidence intervals and in significance testing.

Standard Error of Measurement

In test theory, the standard error of measurement is the standard deviation of observed test scores for a given true score. It is usually estimated with the following formula in which s is the standard deviation of the test scores and r is the reliability of the test.

$$S_{measurement} = s\sqrt{1-r} \tag{14}$$

Standard Error of the Mean

The standard error of the mean is the standard deviation of the sampling distribution of the mean. The formula for the standard error of the mean in a population is:

$$\sigma_m = \frac{\sigma}{\sqrt{N}} \tag{15}$$

where σ is the standard deviation and N is the sample size. When computed in a sample, the estimate of the standard error of the mean is:

$$s_m = \frac{s}{\sqrt{N}} \tag{16}$$

Standard Normal Distribution

The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1.

Standard Score / Standardize / Standard Normal Deviate / Z score

The number of standard deviations a score is from the mean of its population. The term "standard score" is usually used for normal populations; the terms "Z score" and "normal deviate" should only be used in reference to normal distributions. The transformation from a raw score X to a Z score can be done using the following formula:

$$Z = \frac{X - \mu}{\sigma} \tag{17}$$

Transforming a variable in this way is called "standardizing" the variable. It should be kept in mind that if X is not normally distributed then the transformed variable will not be normally distributed either.

Standardize / Standard Score

A variable is standardized if it has a mean of 0 and a standard deviation of 1. The transformation from a raw score X to a standard score can be done using the following formula:

$$X_{standardized} = \frac{X - \mu}{\sigma} \tag{18}$$

where μ is the mean and σ is the standard deviation. Transforming a variable in this way is called "standardizing" the variable. It should be kept in mind that if X is not normally distributed then the transformed variable will not be normally distributed either.

Statistics / Statistic

1. What you are studying right now, also known as statistical analysis, or statistical inference. It is a field of study concerned with summarizing data, interpreting data, and making decisions based on data.
2. A quantity calculated in a sample to estimate a value in a population is called a "statistic."

Stem and Leaf Display

A quasi-graphical representation of numerical data. Generally, all but the final digit of each value is a stem, the final digit is the leaf. The stems are placed in a vertical list, with each matched leaf on one side. They can be very useful for visualizing small data sets with no more than two significant digits. An example is shown below. In this example, you multiply the stems by 10 and add the value of the leaf to obtain the numeric value. Thus the maximum number of touchdown passes is $3 \times 10 + 7 = 37$.

Stem and leaf display of the number of touchdown passes:

$$\begin{array}{c|cccccccccccc} 3 & 2 & 3 & 3 & 7 & & & & & & & & \\ 2 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 3 & 8 & 8 & 9 \\ 1 & 2 & 2 & 4 & 4 & 4 & 5 & 6 & 8 & 8 & 8 & 8 & 9 \quad 9 \\ 0 & 6 & 9 & & & & & & & & & & \end{array} \tag{19}$$

Step

One of the components of a box plot, the step is 1.5 times the difference between the upper hinge and the lower hinge.

Stratified Random Sampling

In stratified random sampling, the population is divided into a number of subgroups (or strata). Random samples are then taken from each subgroup with sample sizes proportional to the size of the subgroup in the population. For instance, if a population contained equal numbers of men and women, and the variable of interest is suspected to vary by gender, one might conduct stratified random sampling to insure a representative sample.

Studentized Range Distribution

The studentized range distribution is used to test the difference between the largest and smallest means. It is similar to the t distribution which is used when there are only two means.

Sturgis' Rule

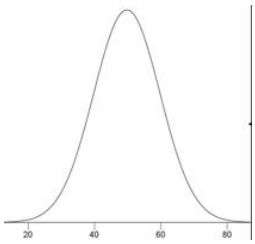
One method of determining the number of classes for a histogram, Sturgis' Rule is to take $1+\log_2(N)$ classes, rounded to the nearest integer.

Sum of Squares Error

In linear regression, the sum of squares error is the sum of squared errors of prediction. In analysis of variance, it is the sum of squared deviations from cell means for between-subjects factors and the $\text{Subjects x Treatment}$ interaction for within-subject factors.

Symmetric Distribution

In a symmetric distribution, the upper and lower halves of the distribution are mirror images of each other. For example, in the distribution shown below, the portions above and below 50 are mirror images of each other. In a symmetric distribution, the mean is equal to the median. Antonym: skewed distribution.



t distribution

The t distribution is the distribution of a value sampled from a normal distribution divided by an estimate of the distribution's standard deviation. In practice the value is typically a statistic such as the mean or the difference between means and the standard deviation is an estimate of the standard error of the statistic. The t distribution in leptokurtic.

t test

Most commonly, a significance test of the difference between means based on the t distribution. Other applications include

a. testing the significance of the difference between a sample mean and a hypothesized value of the mean and
b. testing a specific contrast among means

Third Variable Problem

A type of confounding in which a third variable leads to a mistaken causal relationship between two others. For instance, cities with a greater number of churches have a higher crime rate. However, more churches do not lead to more crime, but instead the third variable, population, leads to both more churches and more crime.

Touchdown Pass

In American football, a touchdown pass occurs when a completed pass results in a touchdown. The pass may be to a player in the end zone or to a player who subsequently runs into the end zone. A touchdown is worth 6 points and allows for a chance at one (and by some rules two) additional point(s).

Trimean

The trimean is a robust measure of central tendency; it is a weighted average of the 25th, 50th, and 75th percentiles. Specifically it is computed as follows:

$$\text{Trimean} = 0.25 \times 25^{th} + 0.5 \times 50^{th} + 0.25 \times 75^{th} \tag{20}$$

True Score

In classical test theory, the true score is a theoretical value that represents a test taker's score without error. If a person took parallel forms of a test thousands of times (assuming no practice or fatigue effects), the mean of all their scores would be a good approximation of their true score since the error would be almost entirely averaged out. It should be distinguished from validity.

Tukey HSD Test

The "Honestly Significantly Different" (HSD) test developed by the statistician John Tukey to test all pairwise comparisons among means. The test is based on the "studentized range distribution."

Two-Tailed Test / Two-Tailed Probability / Non-directional Test

The last step in significance testing involves calculating the probability that a statistic would differ as much or more from the parameter specified in the null hypothesis as does the statistics obtained in the experiment.

A probability computed considering differences in both direction (statistic either larger or smaller than the parameter) is called two-tailed probability. For example, if a parameter is 0 and the statistic is 12, a two-tailed probability would be the he probability of being either ≤ -12 or ≥ 12. Compare with the one-tailed probability which would be the probability of a statistic being $\geq$ to 12 if that were the direction specified in advance.

Type I error

In significance testing, the error of rejecting a true null hypothesis.

Type II error

In significance testing, the failure to reject a false null hypothesis.

Unbiased

A sample is said to be unbiased when every individual has an equal chance of being chosen from the population.

An estimator is unbiased if it does not systematically overestimate or underestimate the parameter it is estimating. In other words, it is unbiased if the mean of the sampling distribution of the statistic is the parameter it is estimating, The sample mean is an unbiased estimate of the population mean.

Unplanned Comparisons / Post Hoc Comparisons

When the comparison among means is decided on after viewing the data, the comparison is called an "unplanned comparison" or a post-hoc comparison. Different statistical tests are required for unplanned comparisons than for planned comparisons.

Upper Adjacent Value

One of the components of a box plot, the higher adjacent value is the largest value in the data below the 75th percentile.

Upper hinge

The upper hinge is one of the components of a box plot; it is the 75th percentile.

Variability / Spread

Variability refers to the extent to which values differ from one another. That is, how much they vary. Variability can also be thought of as how spread out a distribution is. The standard deviation and the semi-interquartile range are measures of variability.

Variable

Something that can take on different values. For example, different subjects in an experiment weigh different amounts. Therefore "weight" is a variable in the experiment. Or, subjects may be given different doses of a drug. This would make "dosage" a variable. Variables can be dependent or independent, qualitative or quantitative, and continuous or discrete.

Variance

The variance is a widely used measure of variability. It is defined as the mean squared deviation of scores from the mean. The formula for variance computed in an entire population is

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \tag{21}$$

where σ^2 represents the variance, μ is the mean, and N is the number of scores.

When computed in a sample in order to estimate the variance in the population, the population is

$$s^2 = \frac{\sum(X-M)^2}{N-1} \tag{22}$$

where s^2 is the estimate of variance, M is the sample mean, and N is the number of scores in the sample.

Variance Sum Law

The variance sum law is an expression for the variance of the sum of two variables. If the variables are independent and therefore Pearson's $r = 0$, the following formula represents the variance of the sum and difference of the variables X and Y:

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y \tag{23}$$

Note that you add the variances for both $X+Y$ and $X-Y$.

If X and Y are correlated, then the following formula (which the former is a special case) should be used:

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho\sigma_X\sigma_Y \tag{24}$$

where ρ is the population value of the correlation. In a sample r is used as an estimate of ρ.

Within-Subjects Design

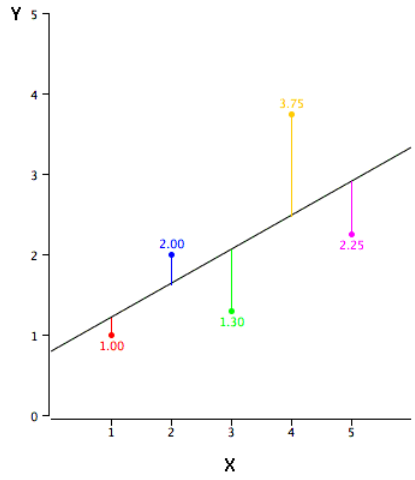
An experimental design in which the independent variable is a within-subjects variable.

Within-Subjects Factor / Within-Subjects Variable / Repeated-Measures Factor / Repeated-Measures Variable

A within-subjects variable is an independent variable that is manipulated by testing each subject at each level of the variable. Compare with a between-subjects variable in which different groups of subjects are used for each level of the variable.

Y-Intercept

The Y-intercept of a line is the value of Y at the point that the line intercepts the Y axis. It is the value of Y when X equals 0. The Y intercept of the black line shown in the graph is 0.785.



Z score / Standard Score / Standardize /Standard Normal Deviate

The number of standard deviations a score is from the mean of its population. The term "standard score" is usually used for normal populations; the terms "Z score" and "normal deviate" should only be used in reference to normal distributions. The transformation from a raw score X to a Z score can be done using the following formula:

$$Z = \frac{X - \mu}{\sigma} \tag{25}$$

Transforming a variable in this way is called "standardizing" the variable. It should be kept in mind that if X is not normally distributed then the transformed variable will not be normally distributed either.

- **Glossary** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Psyc 350:Behavioral Statistics (Toussaint)

Webpages: 213

All licenses found:

- Public Domain: 94.8% (202 pages)
- Undeclared: 5.2% (11 pages)

By Page

- Psyc 350:Behavioral Statistics (Toussaint) - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Introduction to Statistics - *Public Domain*
    - 1.1: What are Statistics? - *Public Domain*
    - 1.2: Importance of Statistics - *Public Domain*
    - 1.3: Descriptive Statistics - *Public Domain*
    - 1.4: Inferential Statistics - *Public Domain*
    - 1.5: Sampling Demonstration - *Public Domain*
    - 1.6: Variables - *Public Domain*
    - 1.7: Percentiles - *Public Domain*
    - 1.8: Levels of Measurement - *Public Domain*
    - 1.9: Measurements - *Public Domain*
    - 1.10: Distributions - *Public Domain*
    - 1.11: Summation Notation - *Public Domain*
    - 1.12: Linear Transformations - *Public Domain*
    - 1.13: Logarithms - *Public Domain*
    - 1.14: Statistical Literacy - *Public Domain*
    - 1.E: Introduction to Statistics (Exercises) - *Public Domain*
  - 2: Graphing Distributions - *Public Domain*
    - 2.1: Graphing Qualitative Variables - *Public Domain*
    - 2.2: Quantitative Variables - *Public Domain*
    - 2.3: Stem and Leaf Displays - *Public Domain*
    - 2.4: Histograms - *Public Domain*
    - 2.5: Frequency Polygons - *Public Domain*
    - 2.6: Box Plots - *Public Domain*
    - 2.7: Box Plot Demo - *Public Domain*
    - 2.8: Bar Charts - *Public Domain*
    - 2.9: Line Graphs - *Public Domain*
    - 2.10: Dot Plots - *Public Domain*
    - 2.11: Statistical Literacy - *Public Domain*
    - 2.E: Graphing Distributions (Exercises) - *Public Domain*
  - 3: Summarizing Distributions - *Public Domain*
    - 3.1: Central Tendency - *Public Domain*
    - 3.2: What is Central Tendency - *Public Domain*
    - 3.3: Measures of Central Tendency - *Public Domain*
    - 3.4: Balance Scale Simulation - *Public Domain*
    - 3.5: Absolute Differences Simulation - *Public Domain*
    - 3.6: Squared Differences Simulation - *Public Domain*
    - 3.7: Median and Mean - *Public Domain*
    - 3.8: Mean and Median Demo - *Public Domain*
    - 3.9: Additional Measures - *Public Domain*
    - 3.10: Comparing Measures - *Public Domain*
    - 3.11: Variability - *Public Domain*
    - 3.12: Measures of Variability - *Public Domain*
    - 3.13: Variability Demo - *Public Domain*
    - 3.14: Estimating Variance Simulation - *Public Domain*
    - 3.15: Shapes of Distributions - *Public Domain*
    - 3.16: Comparing Distributions Demo - *Public Domain*
    - 3.17: Effects of Linear Transformations - *Public Domain*
    - 3.18: Statistical Literacy - *Public Domain*
    - 3.E: Summarizing Distributions (Exercises) - *Public Domain*
  - 4: Normal Distribution - *Public Domain*
    - 4.1: Introduction to Normal Distributions - *Public Domain*
    - 4.2: Areas Under Normal Distributions - *Public Domain*
    - 4.3: Varieties Demonstration - *Public Domain*
    - 4.4: Standard Normal Distribution - *Public Domain*
    - 4.5: Normal Approximation Demonstration - *Public Domain*
    - 4.6: Statistical Literacy - *Public Domain*
    - 4.E: Normal Distribution (Exercises) - *Public Domain*
  - 5: Sampling Distributions - *Public Domain*
    - 5.1: Introduction to Sampling Distributions - *Public Domain*