Introduction to Statistics

More information can be found at Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/)

Primary author and editor:
David M. Lane[1]

Other authors:
David Scott[1], Mikki Hebl[1], Rudy Guerra[1], Dan Osherson[1], and Heidi Zimmer[2]

[1]Rice University; [2]University of Houston, Downtown Campus

Section authors specified on each section.

1. Introduction

2. Graphing Distributions

3. Summarizing Distributions

4. Describing Bivariate Data

1. Introduction

This 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. The only theme common to all of these sections is that they cover concepts and ideas important for other chapters in the book.

- What are Statistics?
- Importance of Statistics
- Descriptive Statistics
- Inferential Statistics
- Variables
- Percentiles
- Measurement
- Levels of Measurement
- Distributions
- Linear Transformations

What Are Statistics

by Mikki Hebl

Learning Objectives

1. Describe the range of applications of statistics
2. Identify situations in which statistics can be misleading
3. 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 will discuss in more detail 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.

Importance of Statistics

by Mikki Hebl

Learning Objectives

1. Give examples of statistics encountered in everyday life
2. 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 electronic 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.

Now let us get to work!

Descriptive Statistics

by Mikki Hebl

Prerequisites

• none

Learning Objectives

1. Define "descriptive statistics"
2. 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, an 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. The table shows the average salaries for various occupations in the United States in 1999.

Table 1. Average salaries for various occupations in 1999.

$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 2. It 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 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
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.) The following table shows the winning times for both men and women (the latter have only been allowed to compete since 1984).

Table 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
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 far 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.

Inferential Statistics

by Mikki Hebl

Prerequisites
• Chapter 1: Descriptive Statistics

Learning Objectives
1. Distinguish between a sample and a population
2. Define inferential statistics
3. Identify biased samples
4. Distinguish between simple random sampling and stratified sampling
5. 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: 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. Who 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 #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 #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 #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 #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 #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 can't 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).

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 #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 #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 in Chapter 5.) 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.

Variables

by Heidi Ziemer

Prerequisites

- none

Learning Objectives

1. Define and distinguish between independent and dependent variables
2. Define and distinguish between discrete and continuous variables
3. 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: 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.

1. What is the independent variable? (dietary supplement: none, blueberry, strawberry, and spinach)

2. What are the dependent variables? (memory test and motor skills test)

Example #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.

1. What is the independent variable? (supplements: beta-carotene or placebo)

2. What is the dependent variable? (occurrence of cancer)

Example #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.

1. What is the independent variable? (brightness of brake lights)

2. What is the dependent variable? (time to hit brakes)

Levels of an Independent Variable

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 5 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 previously, 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 1.64237123922121 seconds. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

Percentiles

by David Lane

Prerequisites

- none

Learning Objectives

1. Define percentiles
2. 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.

A 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. Notice the numbers are given ranks ranging from 1 for the lowest number to 8 for the highest number.

Table 1. Test Scores.

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

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

$$R = \frac{P}{100} \times (N + 1)$$

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

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

If R is an integer, the Pth percentile is be 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 I_R and with Rank $I_R + 1$. 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 F_R 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 2.

Table 2. 20 Quiz Scores.

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

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

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

IR = 5 and FR = 0.25.

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:

25th percentile = (.25) x (5 - 5) + 5 = 5.

For the 85th percentile,

$$R = \frac{85}{100} \times (20 + 1) = 17.85$$

IR = 17 and FR = 0.85

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

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

$$R = \frac{50}{100} \times (4 + 1) = 2.5$$

IR = 2 and FR = 0.5.

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.

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

$$R = \frac{50}{100} \times (5 + 1) = 3$$

IR = 3 and FR = 0.

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:

50th percentile = (0.00) (9 - 5) + 5 = 5.

Levels of Measurement

by Dan Osherson and David M. Lane

Prerequisites

- Chapter 1: Variables

Learning Objectives

1. Define and distinguish among nominal, ordinal, interval, and ratio scales
2. Identify a scale type
3. Discuss the type of scale used in psychological measurement
4. 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.

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 (i) the difference between Subject A's score of 2 and Subject B's score of 3 and (ii) 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:

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:

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. 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, there are extreme situations in which computing the mean of an ordinally-measured variable can be very misleading.

Distributions

by David M. Lane and Heidi Ziemer

Prerequisites

- Chapter 1: Variables

Learning Objectives

1. Define "distribution"
2. Interpret a frequency distribution
3. Distinguish between a frequency distribution and a probability distribution
4. Construct a grouped frequency distribution for a continuous variable
5. Identify the skew of a distribution
6. Identify bimodal, leptokurtic, and platykurtic distributions

Distributions of Discrete Variables

I recently purchased a bag of Plain M&M's. The M&M's were in 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.

Table 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.

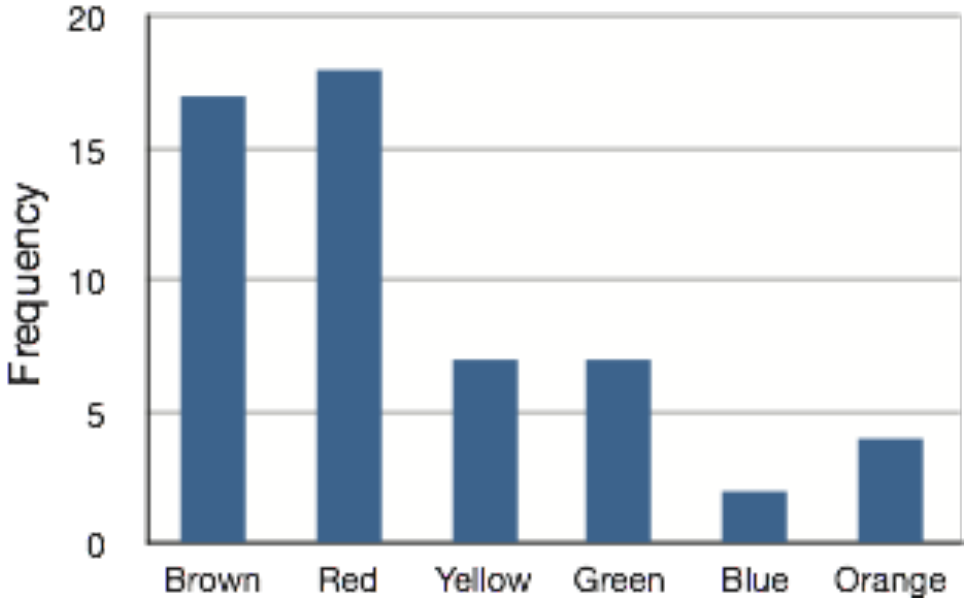


Figure 1. Distribution of 55 M&M's.

The distribution shown in Figure 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 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 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).

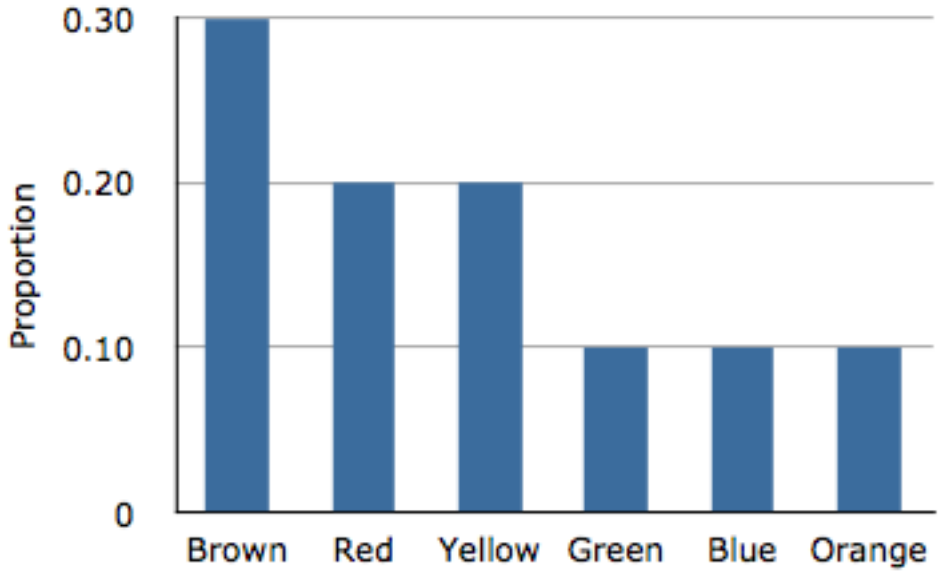


Figure 2. Distribution of all M&M's.

Notice that the distributions in Figures 1 and 2 are not identical. Figure 1 portrays the distribution in a sample of 55 M&M's. Figure 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 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 2 are the times it took one of us (DL) to move the cursor 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 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 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 3 shows a grouped frequency distribution for these 20 times.

Table 3. Grouped frequency distribution

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

Grouped frequency distributions can be portrayed graphically. Figure 3 shows a graphical representation of the frequency distribution in Table 3. This kind of graph is called a histogram. Chapter 2 contains an entire section devoted to histograms.

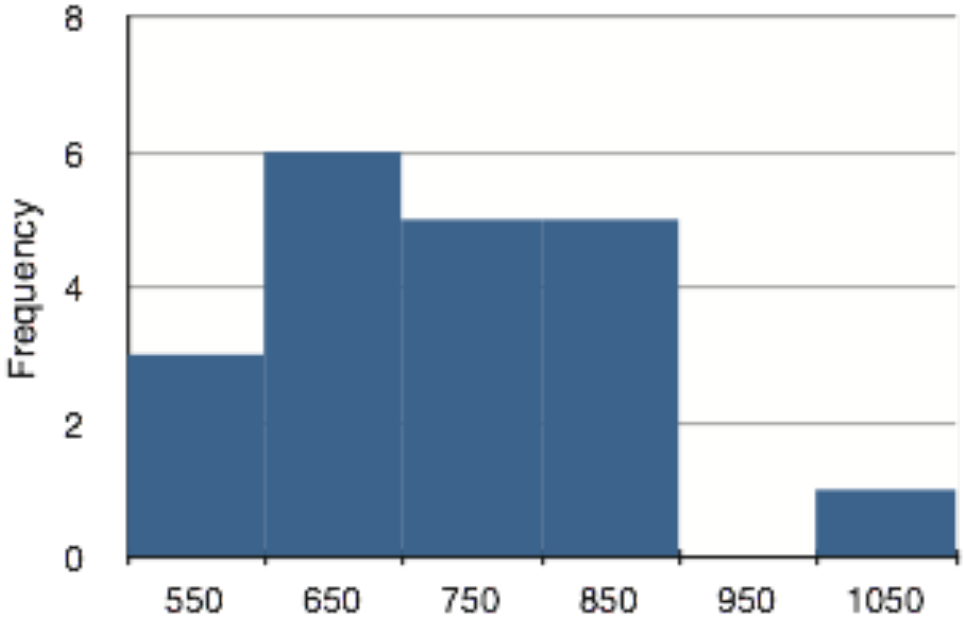


Figure 3. A histogram of the grouped frequency distribution shown in Table 3. The labels on the X-axis are the middle values of the range they represent.

Probability Densities

The histogram in Figure 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 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 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, see Chapter 7.)

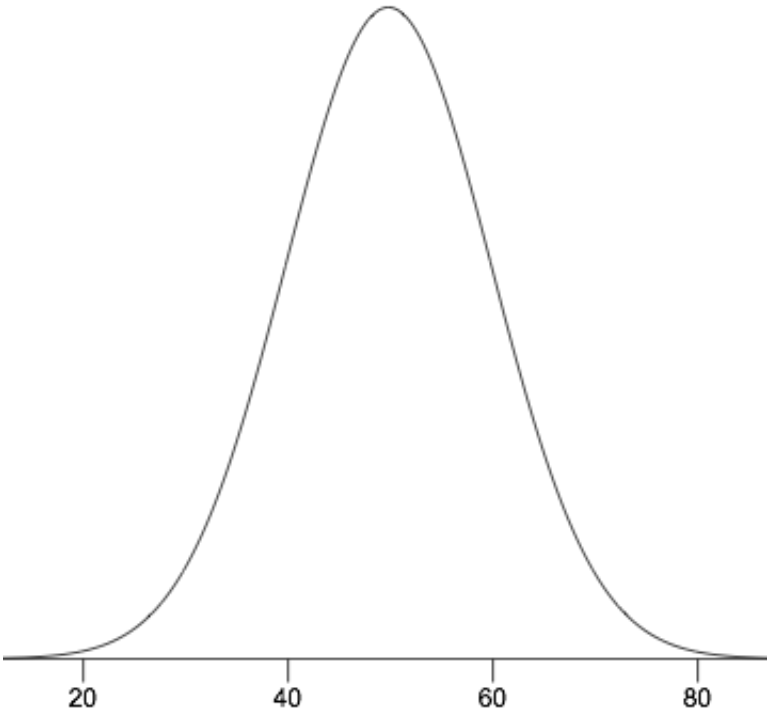


Figure 4. A normal distribution.

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.

Shapes of Distributions

Distributions have different shapes; they don't all look like the normal distribution in Figure 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 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 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 4 is symmetric; if you folded it in the middle, the two sides would match perfectly. Figure 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.”

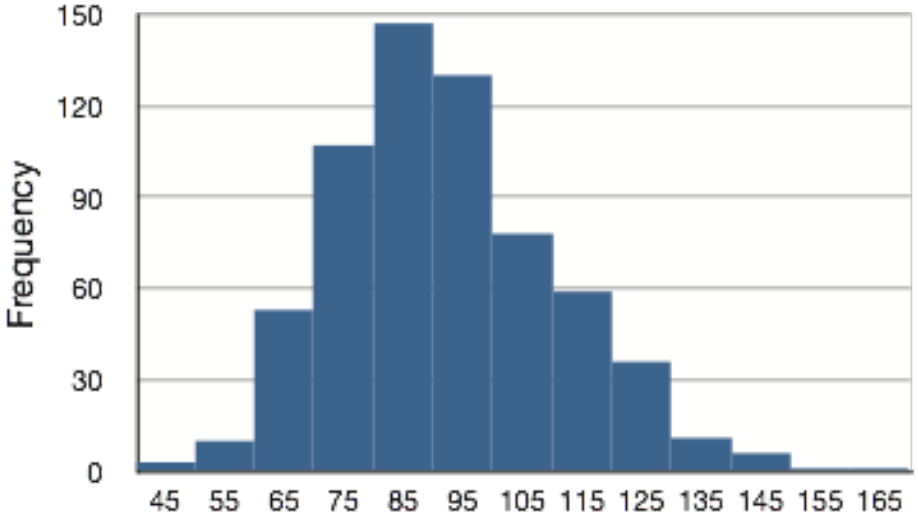


Figure 5. A distribution with a positive skew.

Figure 6 shows the salaries of major league baseball players in 1974 (in thousands of dollars). This distribution has an extreme positive skew.

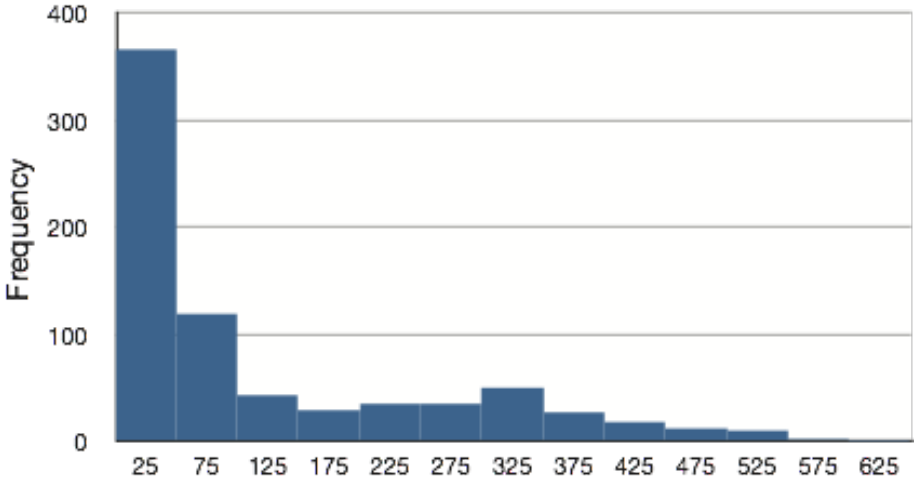


Figure 6. A distribution with a very large positive skew.

A continuous distribution with a positive skew is shown in Figure 7.

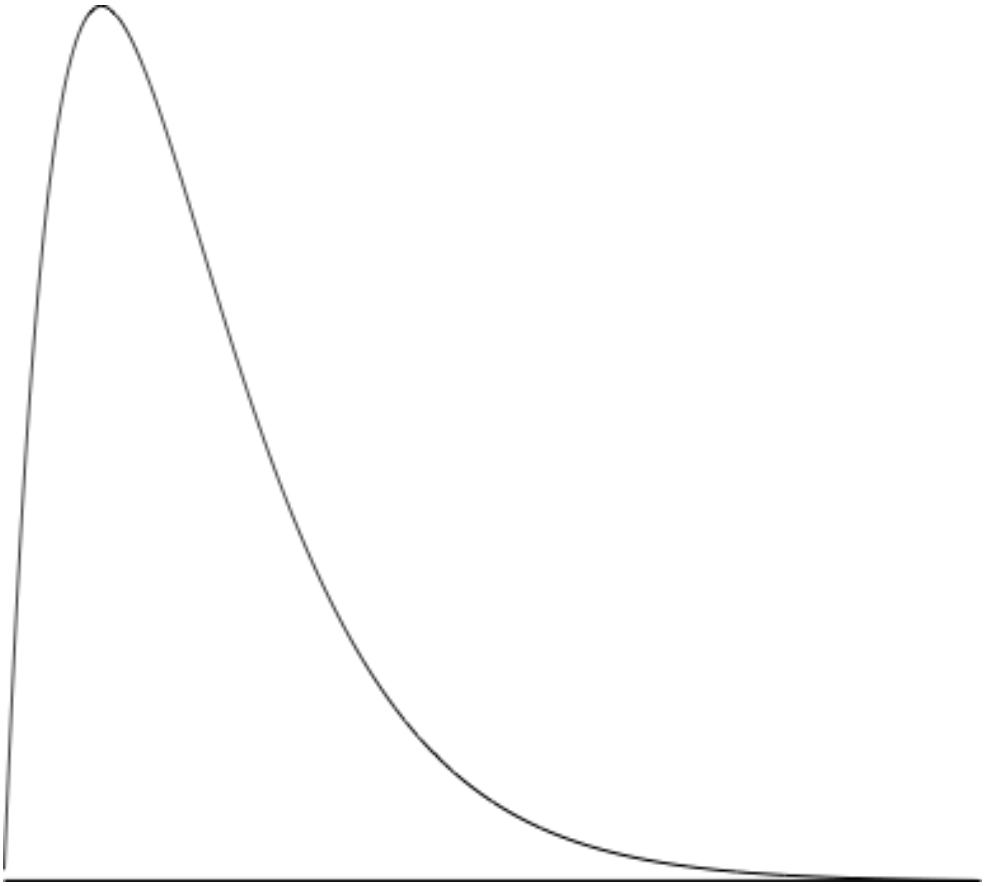


Figure 7. A continuous distribution with a positive skew.

Although less common, some distributions have a negative skew. Figure 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.

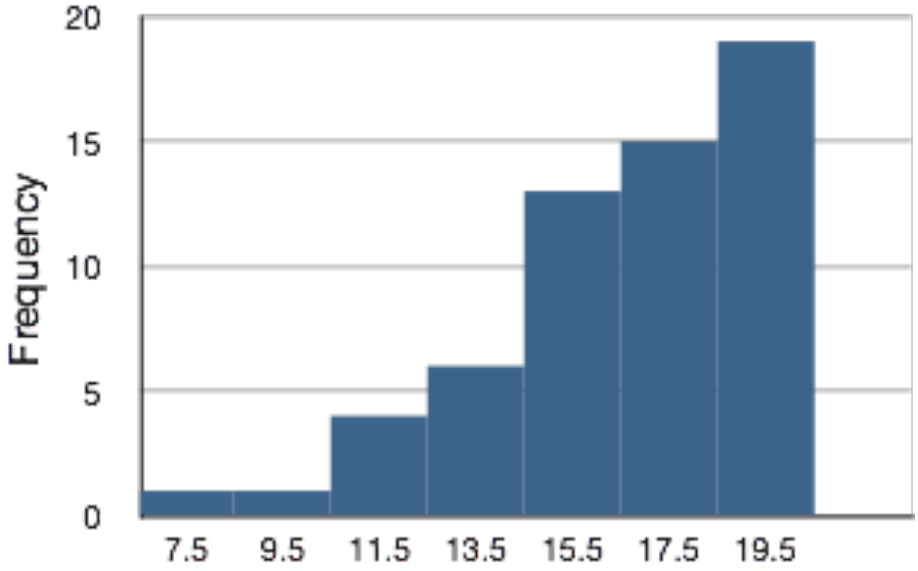


Figure 8. A distribution with negative skew. This histogram shows the frequencies of various scores on a 20-point question on a statistics test.

A continuous distribution with a negative skew is shown in Figure 9.

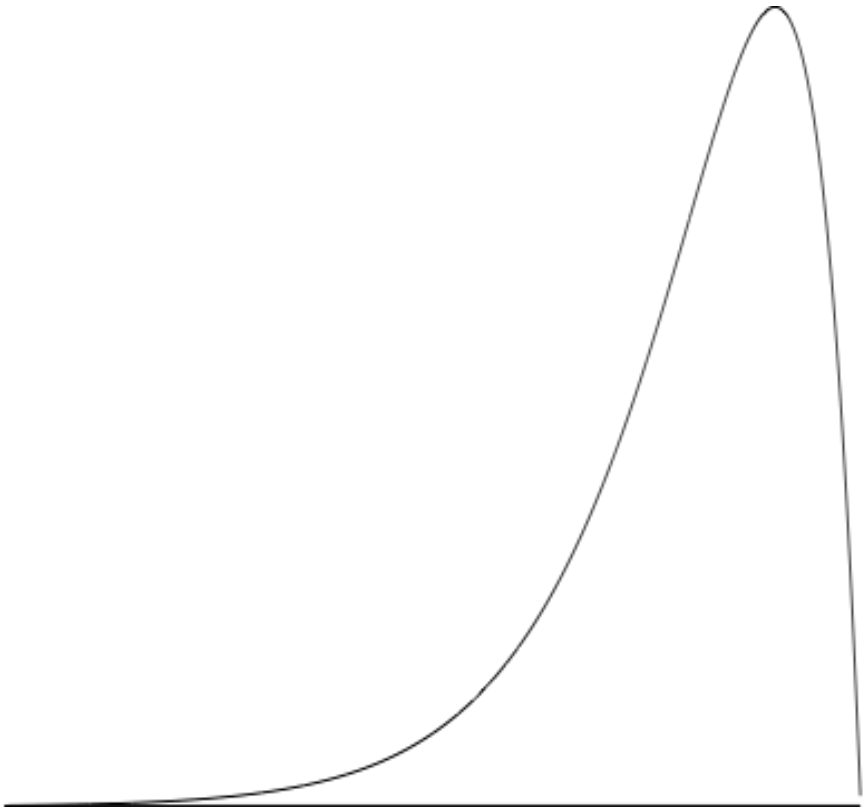


Figure 9. A continuous distribution with a negative skew.

The distributions shown so far all have one distinct high point or peak. The distribution in Figure 10 has two distinct peaks. A distribution with two peaks is called a bimodal distribution.

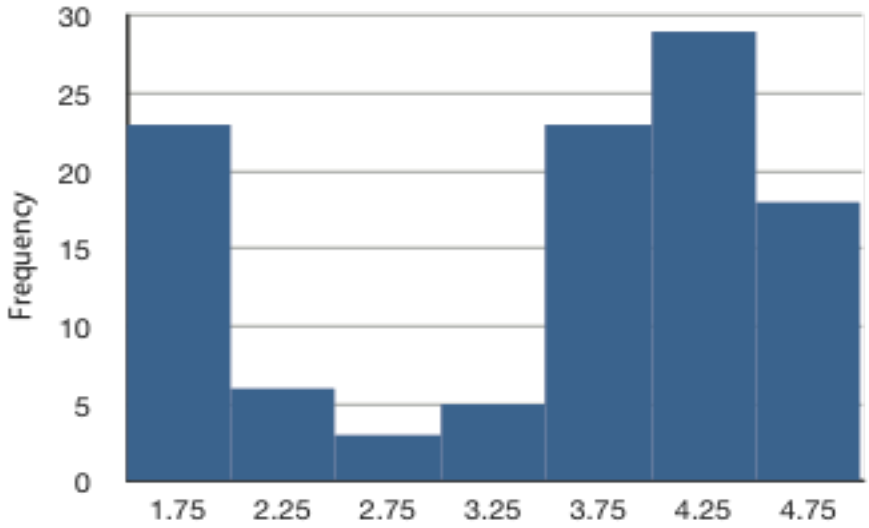


Figure 10. Frequencies of times between eruptions of the Old Faithful geyser. Notice the two distinct peaks: one at 1.75 and the other at 4.25.

Distributions also differ from each other in terms of how large or "fat" their tails are. Figure 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.

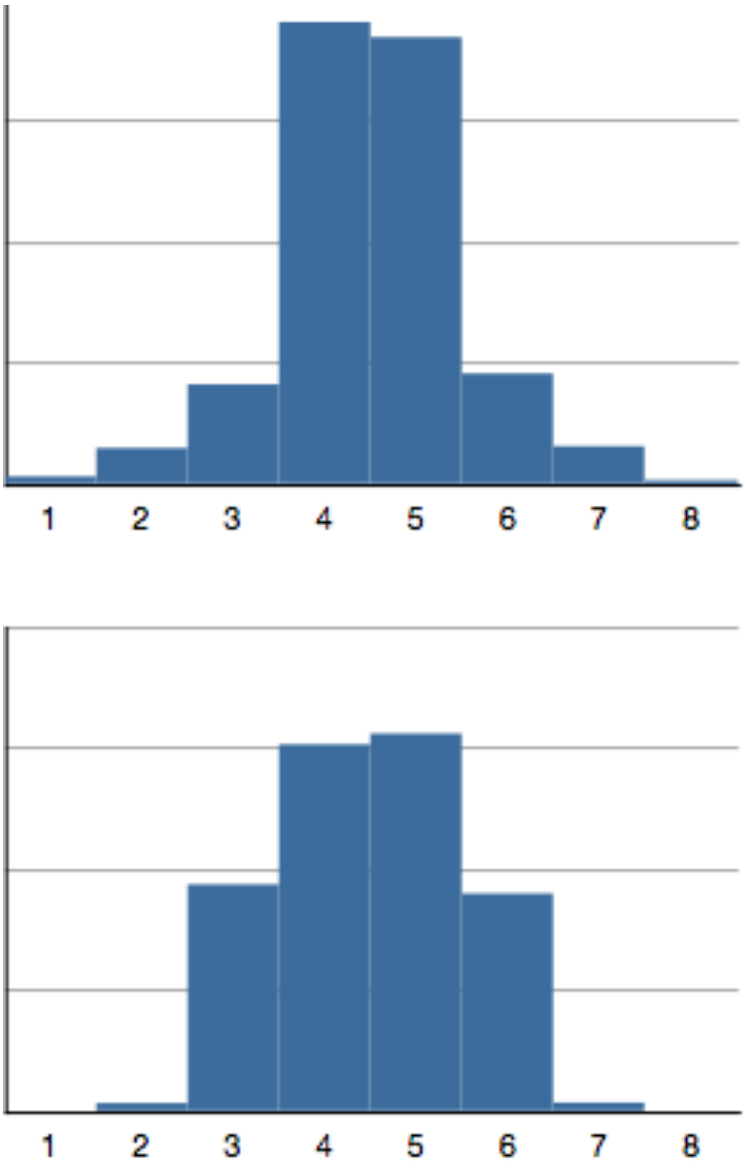


Figure 11. Distributions differing in kurtosis. The top distribution has long tails. It is called “leptokurtic.” The bottom distribution has short tails. It is called “platykurtic.”

Linear Transformations

by David M. Lane

Prerequisites
- None

Learning Objectives
1. Give the formula for a linear transformation
2. Determine whether a transformation is linear
3. 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 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. 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 Centigrade and degrees Fahrenheit. Table 2 shows the temperatures of 5 US cities in the early afternoon of November 16, 2002.

Table 2. 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

The formula to transform Centigrade to Fahrenheit is:

```
F = 1.8C + 32
```

The formula for converting from Fahrenheit to Centigrade is

```
C = 0.5556F - 17.778
```

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

Figure 1 shows a plot of degrees Centigrade 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.

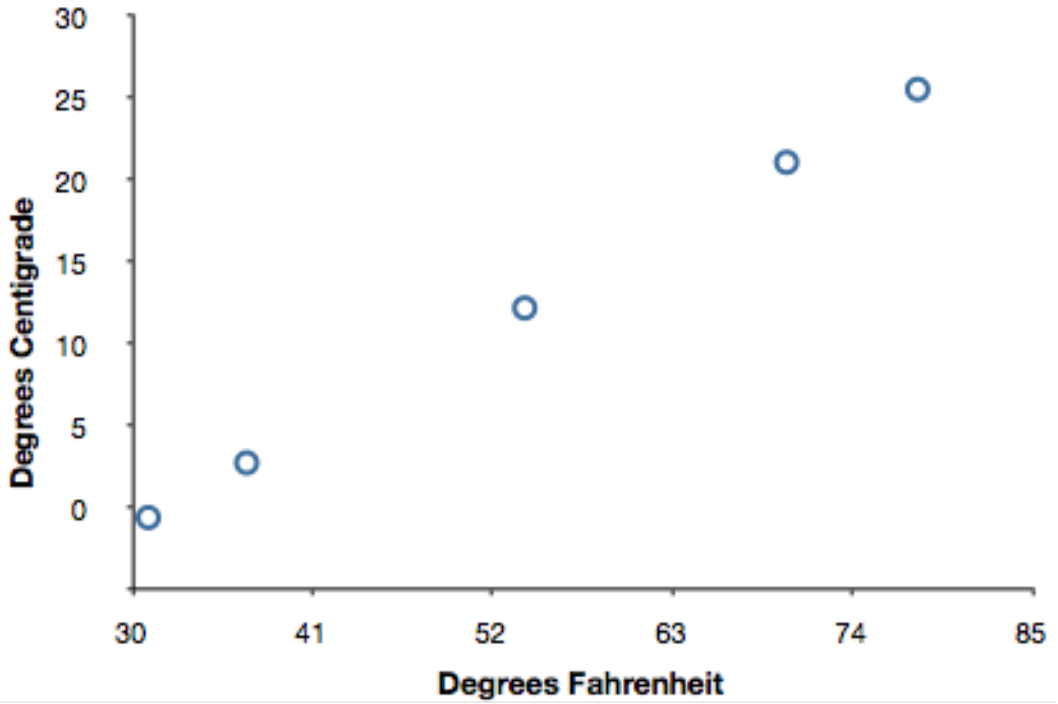


Figure 1. Degrees Centigrade as a function of degrees Fahrenheit

2. Graphing Distributions

A. Qualitative Variables
B. Quantitative Variables
    1. Stem and Leaf Displays
    2. Histograms
    3. Frequency Polygons
    4. Box Plots
    5. Bar Charts
    6. Line Graphs
    7. Dot Plots
C. Exercises

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 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.

by David M. Lane

Prerequisites
- Chapter 1: Variables

Learning Objectives
1. Create a frequency table
2. Determine when pie charts are valuable and when they are not
3. Create and interpret bar charts
4. 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 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" of 0.17 = 85/500.

Table 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 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.

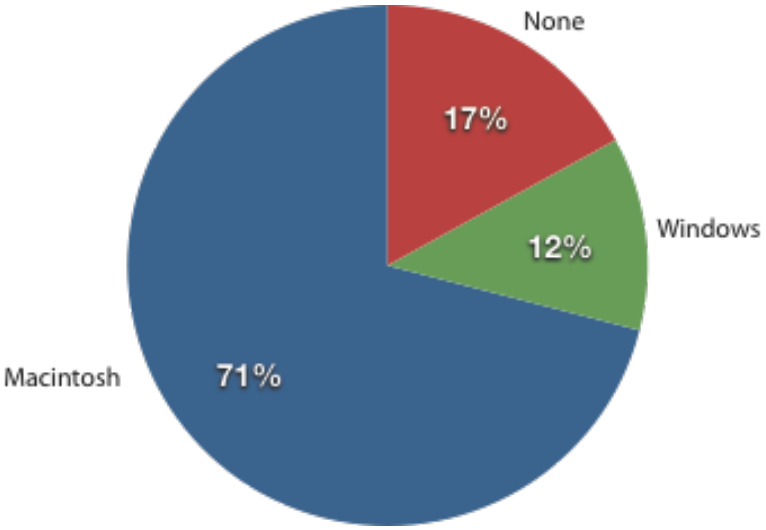


Figure 1. Pie chart of iMac purchases illustrating frequencies of previous computer ownership.

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. 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 in each category as is typical in pie charts.

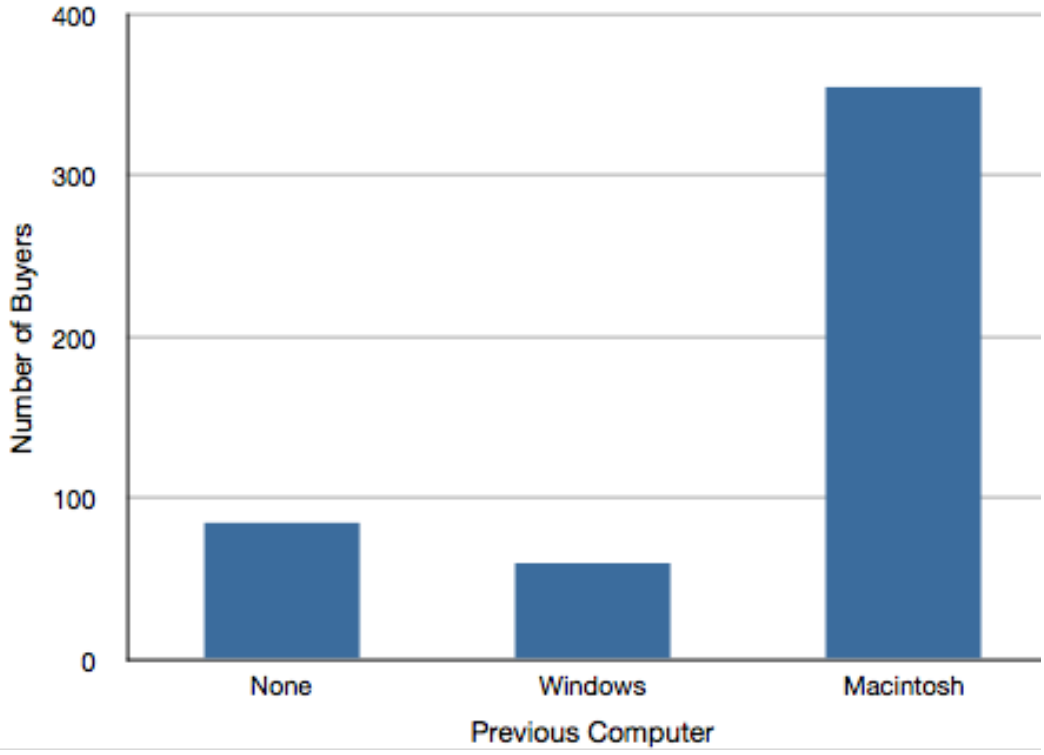


Figure 2. Bar chart of iMac purchases as a function of previous computer ownership.

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 3 shows the number of people playing card games at the Yahoo web site 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.

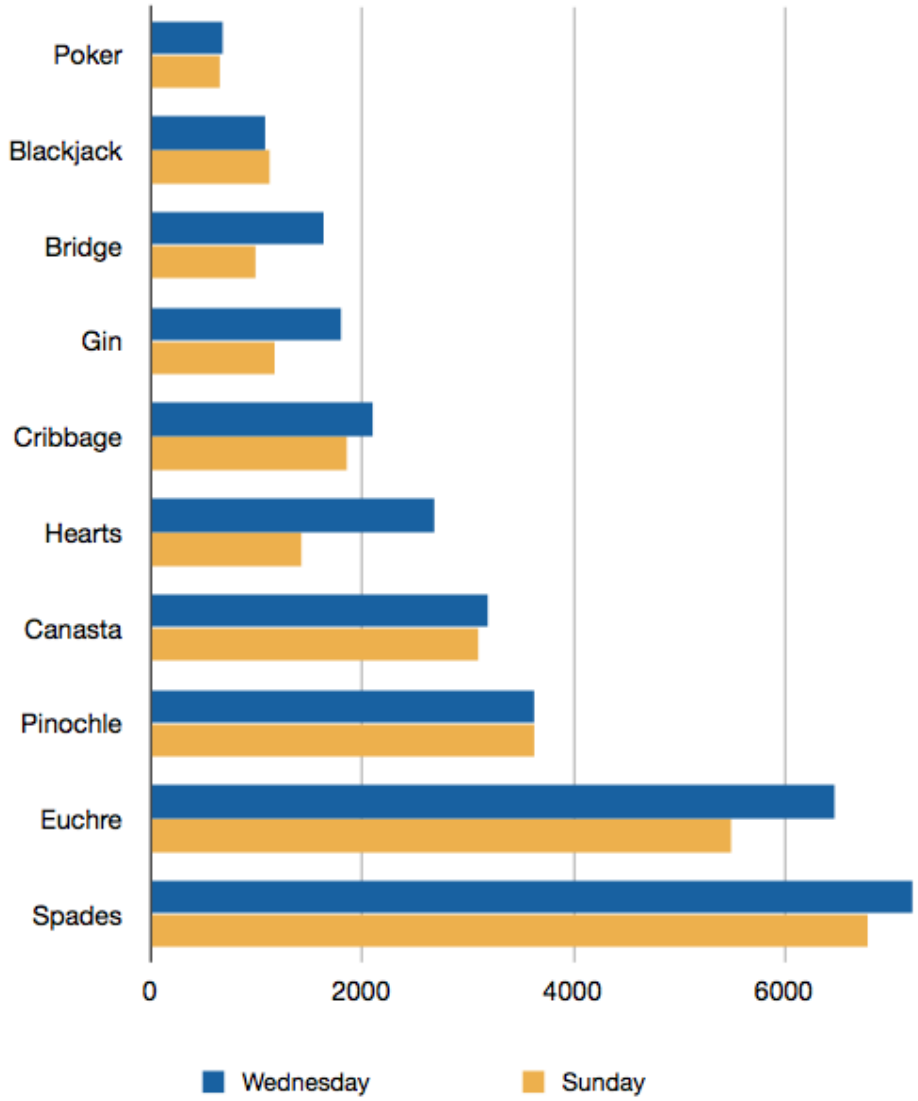


Figure 3. A bar chart of the number of people playing different card games on Sunday and Wednesday.

The bars in Figure 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 this chapter.

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 4 are usually not as effective as their two-dimensional counterparts.

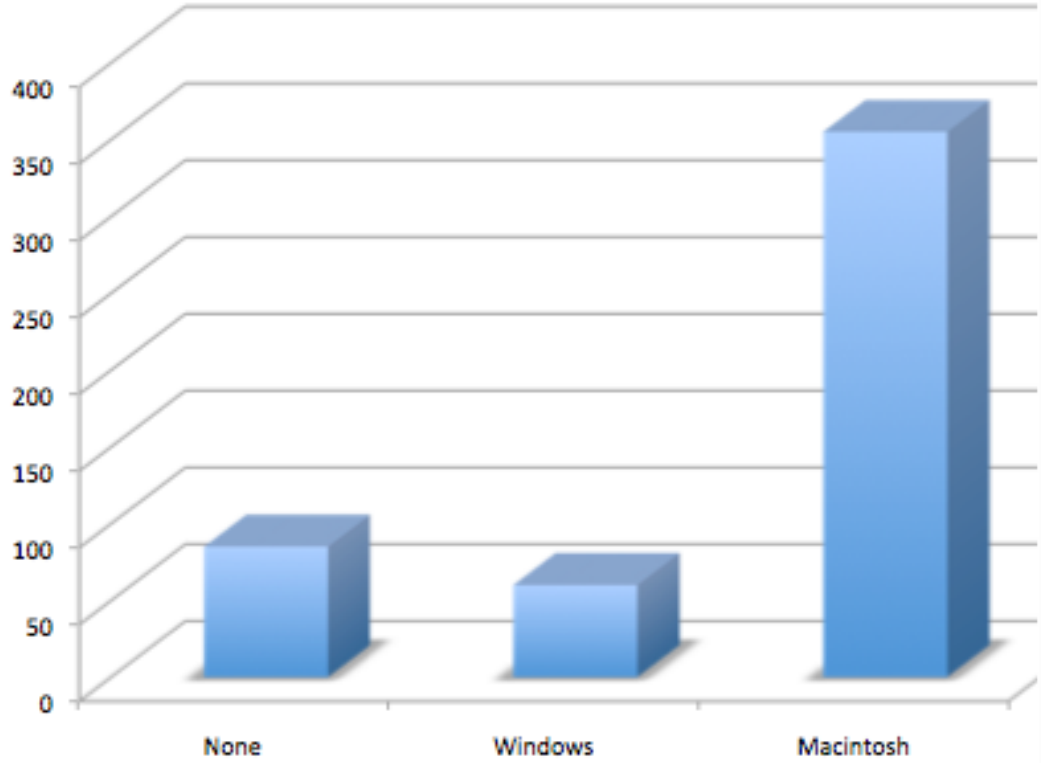


Figure 4. A three-dimensional version of Figure 2.

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure 5 presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 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 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 5 instead of Figure 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.

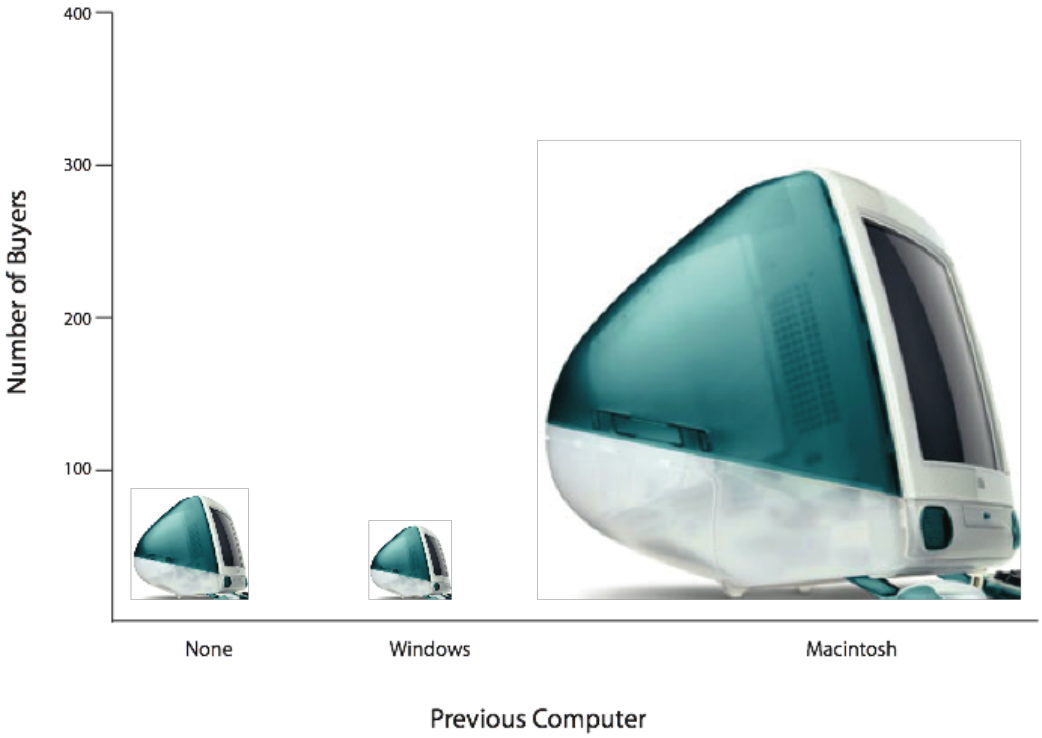


Figure 5. A redrawing of Figure 2 with a lie factor greater than 8.

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 6 shows the iMac data with a baseline of 50. Once again, the differences in areas suggests a different story than the true differences in percentages. The number of Windows-switchers seems minuscule compared to its true value of 12%.

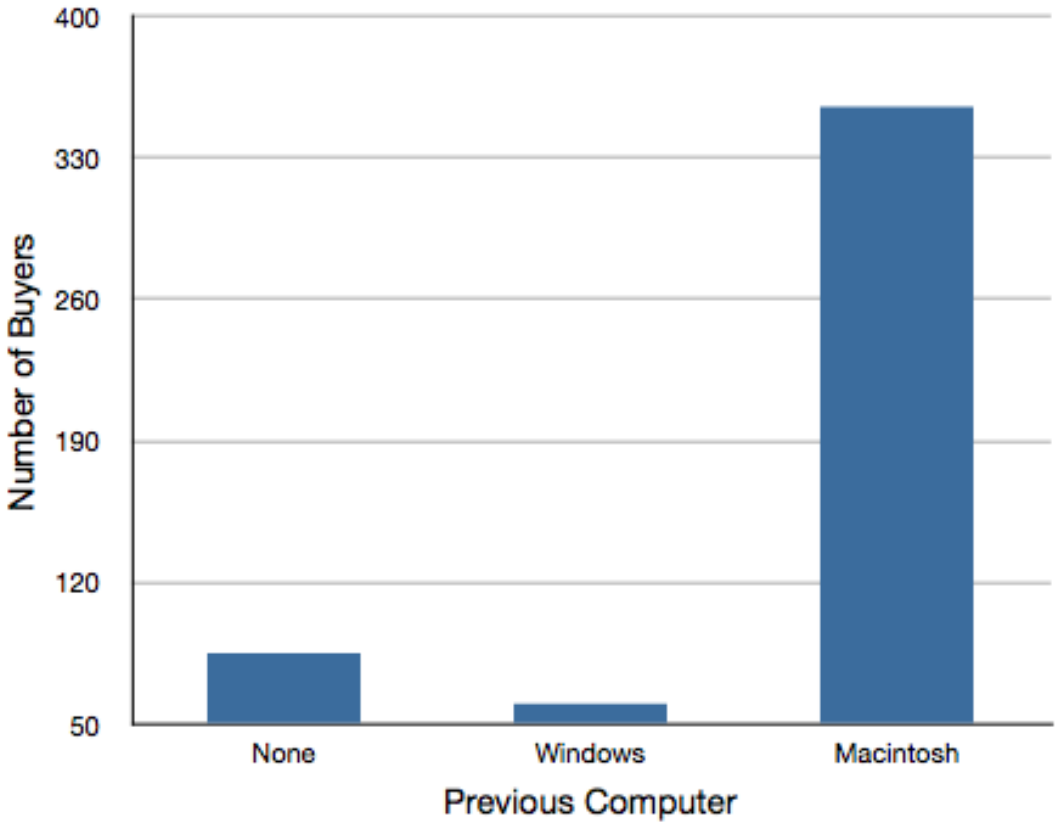


Figure 6. A redrawing of Figure 2 with a baseline of 50.

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 7 inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure 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.

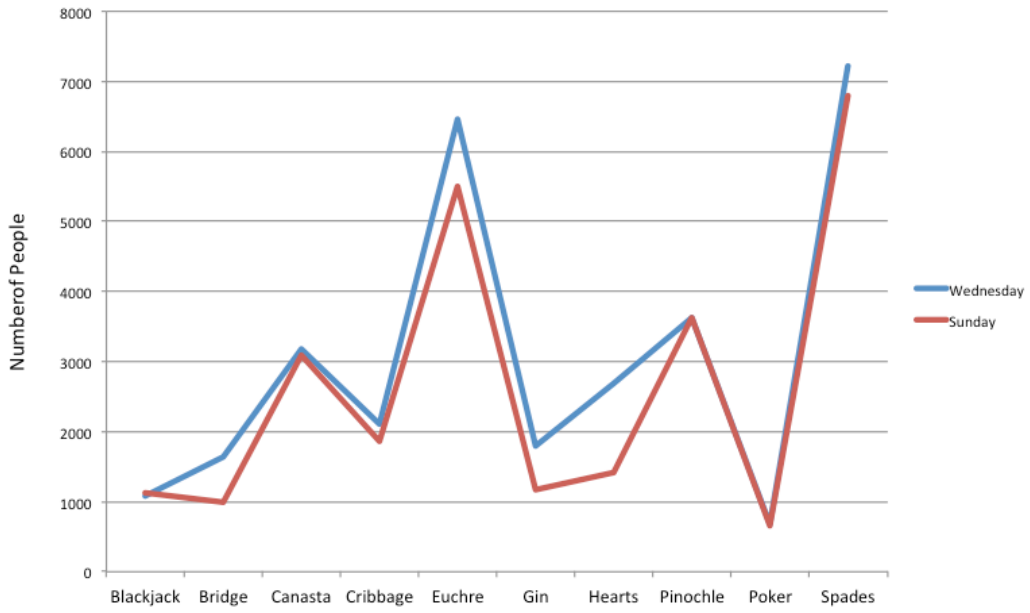


Figure 7. A line graph used inappropriately to depict the number of people playing different card games on Sunday and Wednesday.

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.

Graphing Quantitative Variables

- Stem and Leaf Displays
- Histograms
- Frequency Polygons
- Box Plots
- Bar Charts
- Line Graphs
- Dot Plots

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, 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) dot plots, and (8) scatter plots (discussed in a different chapter). 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.

Stem and Leaf Displays

by David M. Lane

Prerequisites

- Chapter 1: Distributions

Learning Objectives

1. Create and interpret basic stem and leaf displays
2. Create and interpret back-to-back stem and leaf displays
3. 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, we will start with an example. Consider Table 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 1. Number of touchdown passes.

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

A stem and leaf display of the data is shown in Figure 1. The left portion of Figure 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.

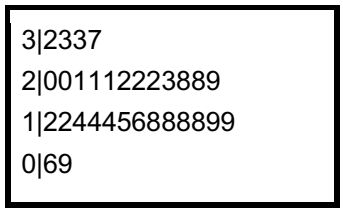


Figure 1. Stem and leaf display of the number of touchdown passes.

To make this clear, let us examine Figure 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 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 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 1 than in Table 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 TD's, 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 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 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.

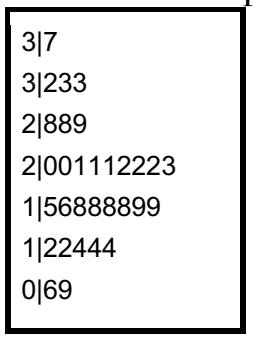


Figure 2. Stem and leaf display with the stems split in two.

Figure 2 is more revealing than Figure 1 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 display." Figure 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.

11	4	
	3	7
332	3	233
8865	2	889
44331110	2	001112223
987776665	1	56888899
321	1	22444
7	0	69

Figure 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 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 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. The effects of priming (thousandths of a second).

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

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 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.

```
 4|33
 3|6
 2|00456
 1|00134
 0|1245589
-0|0679
-1|005559
-2|7
```

Figure 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, 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 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.

```
4|899
4|6
4|4455
4|333
4|01
3|99
3|677777
3|55
3|223
3|111
2|8899
2|666667
2|444455
2|22333
2|000000
1|88888888888999999999
1|666666777777
1|444444444445555555555
1|22222222222222222333333333
1|00000000000000011111111111111111111111111
```

Figure 5. Stem and leaf display of populations of 185 US cities with populations between 100,000 and 500,000 in 1988.

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 display 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 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!

Histograms

by David M. Lane

Prerequisites

- Chapter 1: Distributions
- Chapter 2: Graphing Qualitative Data

Learning Objectives

1. Create a grouped frequency distribution
2. Create a histogram based on a grouped frequency distribution
3. 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 1.

Table 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 1.

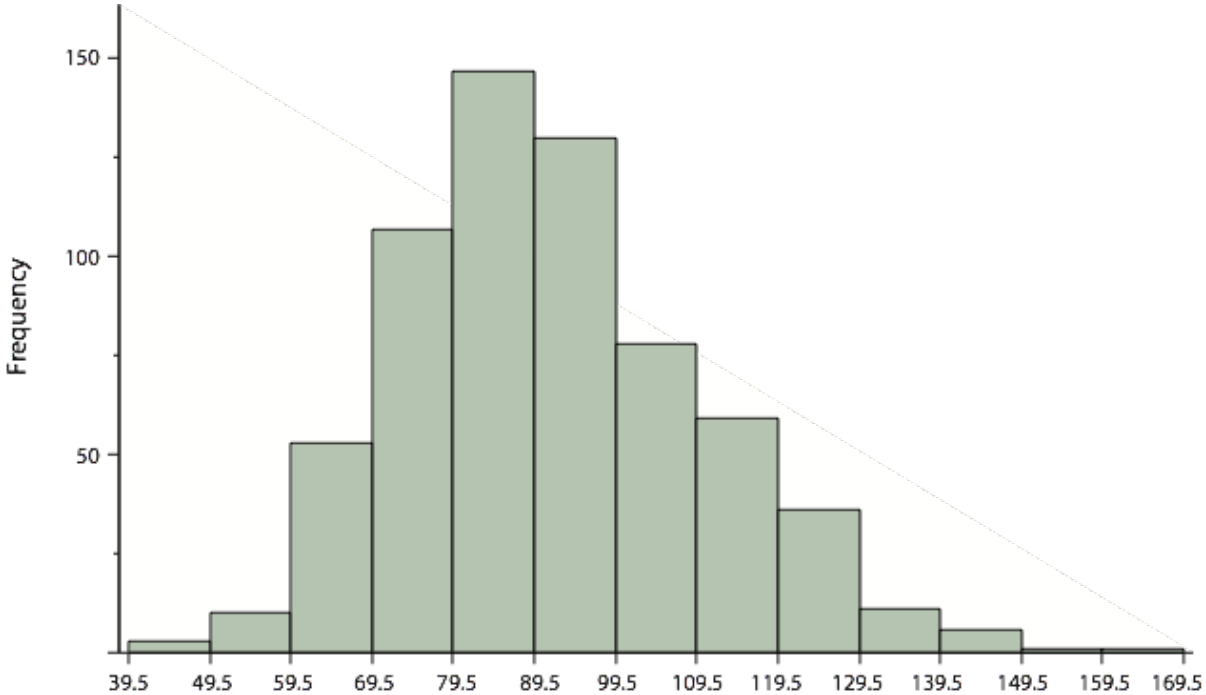


Figure 1. Histogram of scores on a psychology test.

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 Chapter 3.)

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).

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 $1 + \text{Log}_2(N)$, where $\text{Log}_2(N)$ is the base 2 log of the number of observations. The formula can also be written as $1 + 3.3\ \text{Log}_{10}(N)$ where $\text{Log}_{10}(N)$ 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 $\text{Log}_2(1000)$. 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 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 (external link; Java required). The demonstration reveals the consequences of different choices of bin width and of lower boundary for the first interval.

Frequency Polygons

by David M. Lane

Prerequisites

- Chapter 2: Histograms

Learning Objectives

1. Create and interpret frequency polygons
2. Create and interpret cumulative frequency polygons
3. 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.

A frequency polygon for 642 psychology test scores shown in Figure 1 was constructed from the frequency table shown in Table 1.

Table 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
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	170.5	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 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.

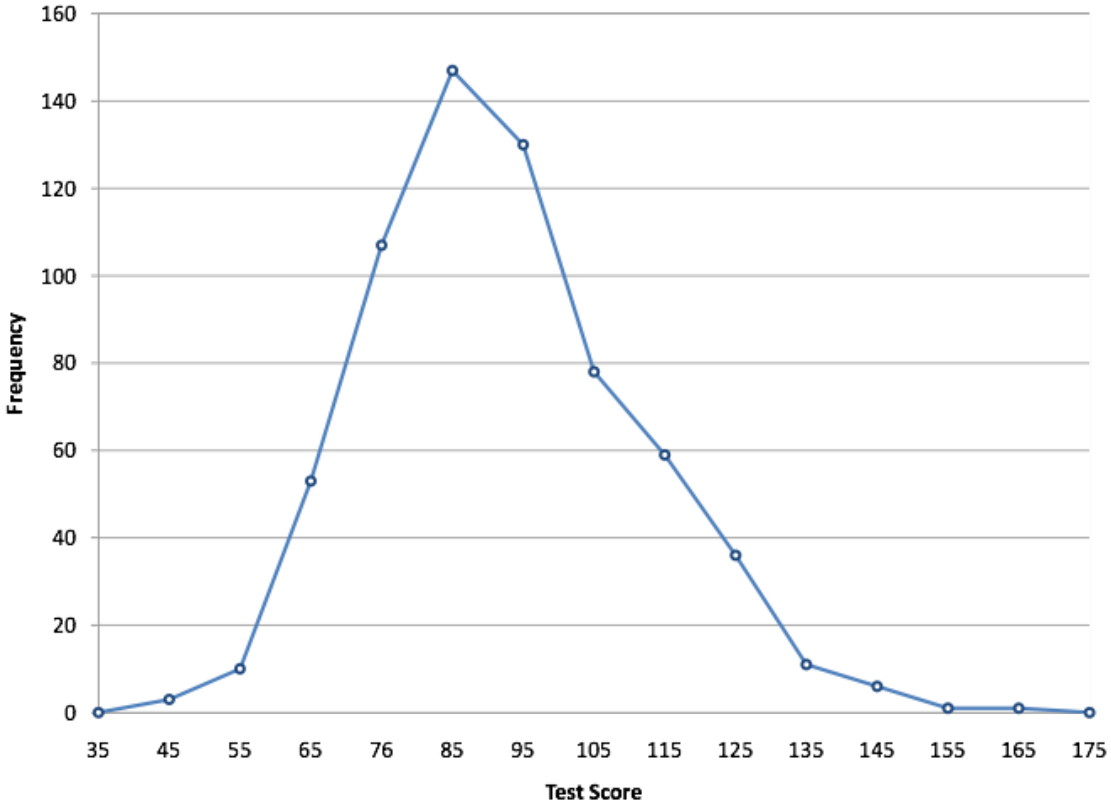


Figure 1. Frequency polygon for the psychology test scores.

A cumulative frequency polygon for the same test scores is shown in Figure 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.

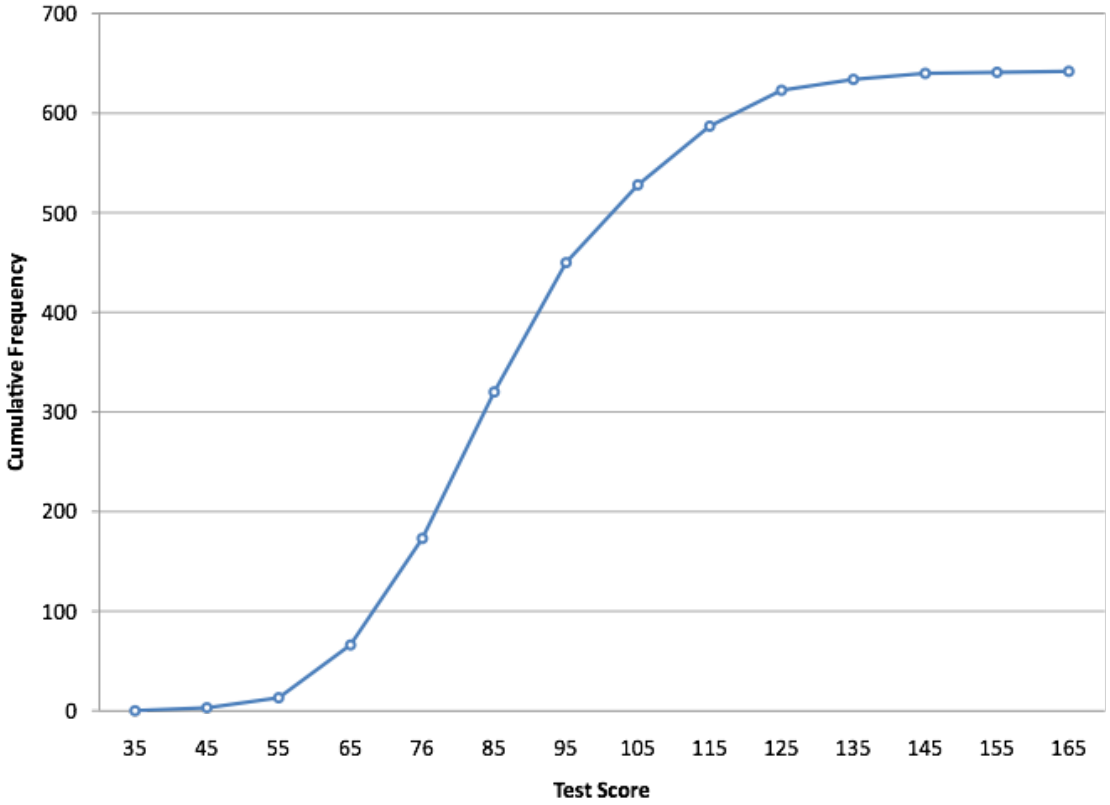


Figure 2. Cumulative frequency polygon for the psychology test scores.

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 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 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.

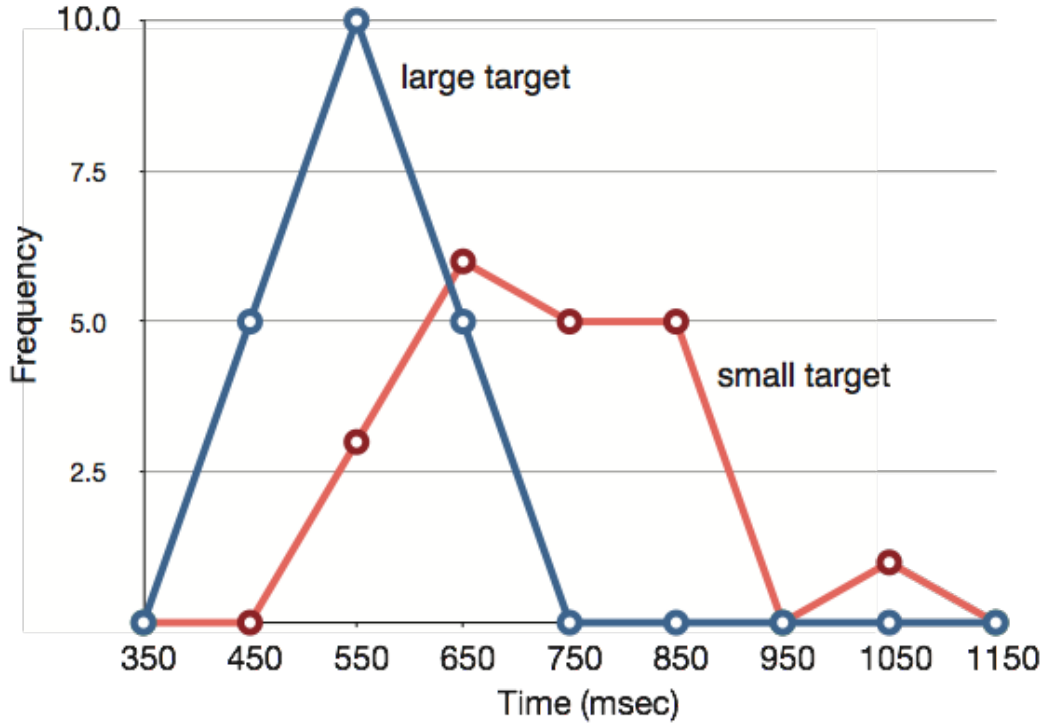


Figure 3. Overlaid frequency polygons.

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in Figure 4 using the same data from the cursor task. The difference in distributions for the two targets is again evident.

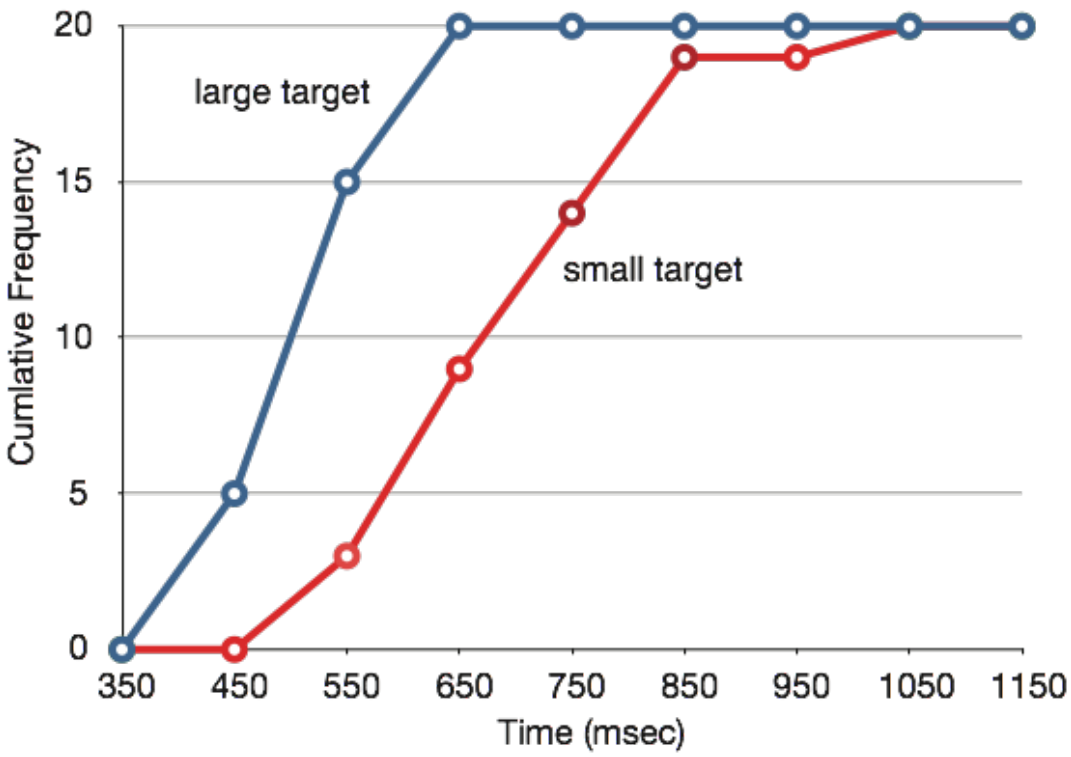


Figure 4. Overlaid cumulative frequency polygons.

Box Plots

by David M. Lane

Prerequisites

- Chapter 1: Percentiles
- Chapter 2: Histograms
- Chapter 2: Frequency Polygons

Learning Objectives

1. Define basic terms including hinges, H-spread, step, adjacent value, outside value, and far out value
2. Create a box plot
3. Create parallel box plots
4. 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. 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 75th percentiles in the distribution of scores. Figure 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 1.

Table 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.

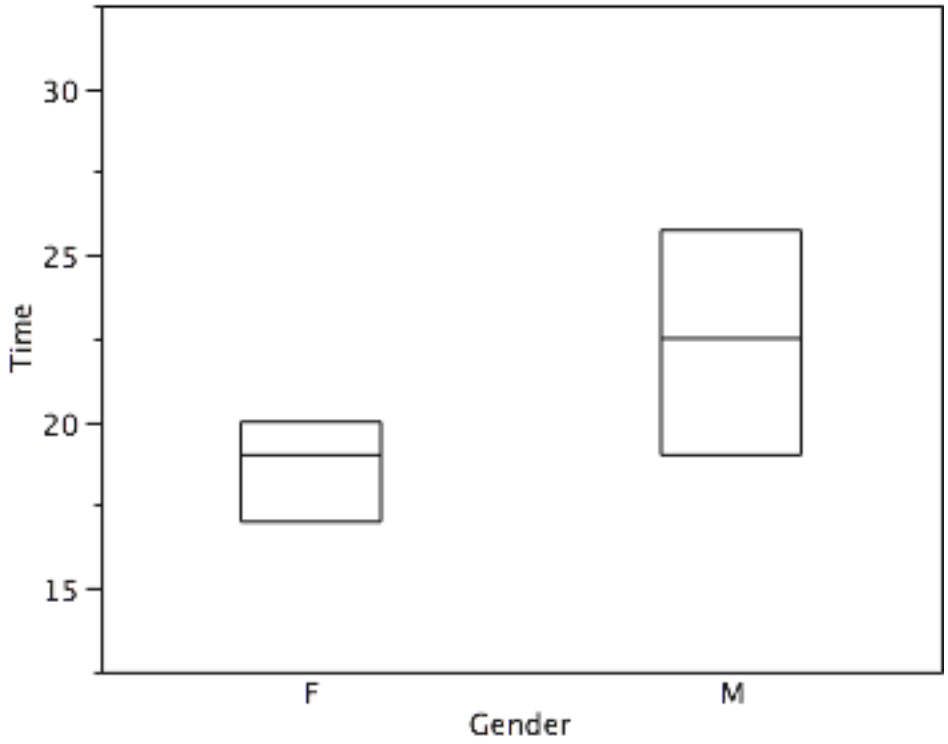


Figure 1. The first step in creating box plots.

Before proceeding, the terminology in Table 2 is helpful.

Table 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 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).

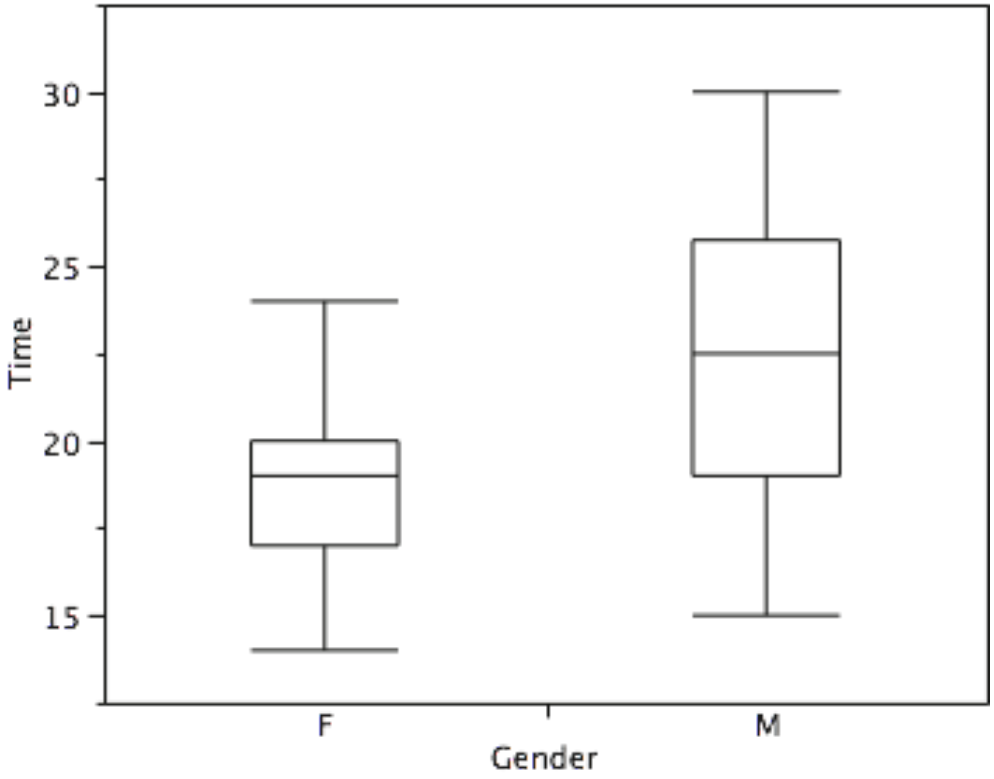


Figure 2. The box plots with the whiskers drawn.

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 3.

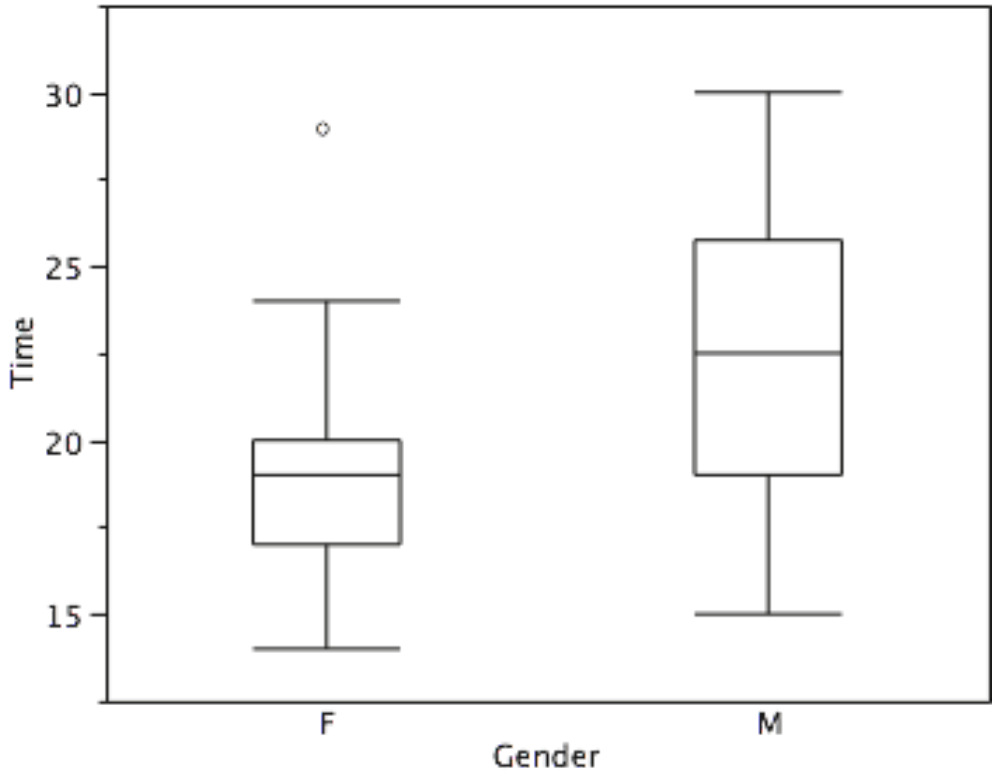


Figure 3. The box plots with the outside value shown.

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 4 shows the result of adding means to our box plots.

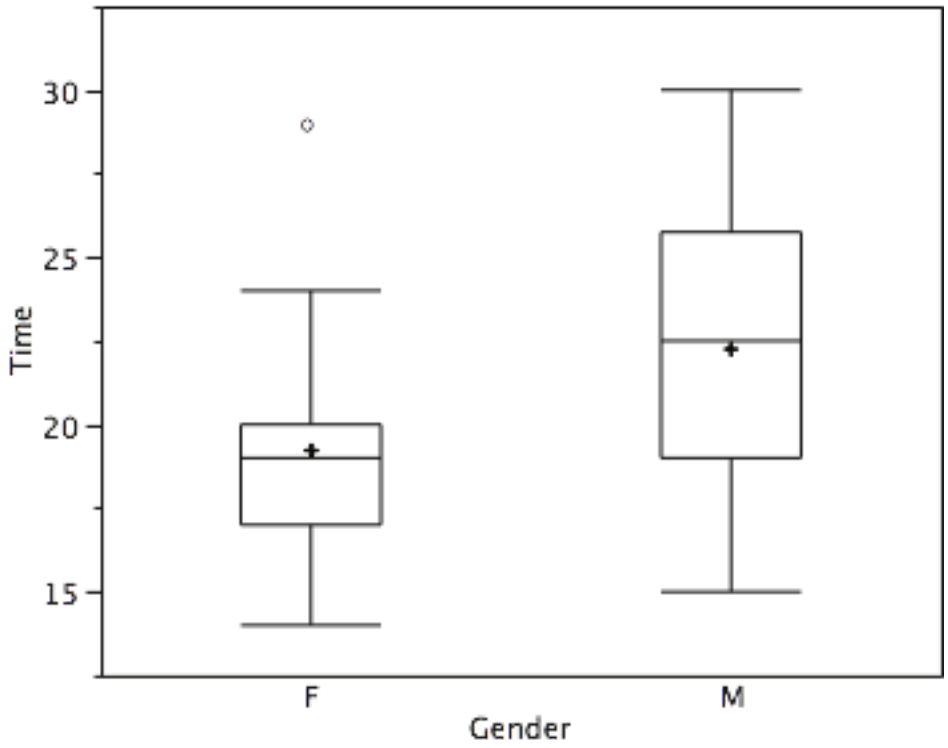


Figure 4. The completed box plots.

Figure 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 seconds. 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 5 shows the box plot for the women's data with detailed labels.

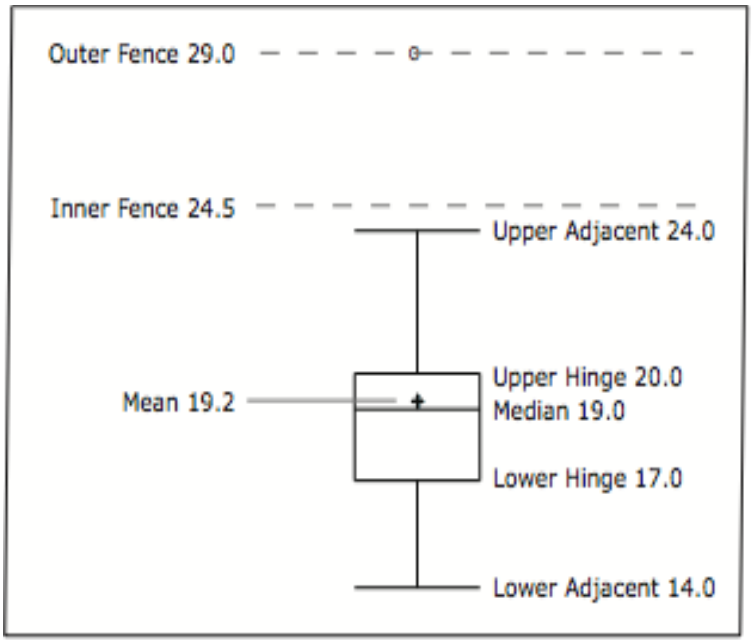


Figure 5. The box plots for the women's data with detailed labels.

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 use create a histogram and/or a stem and leaf display.

Variations on box plots

Statistical analysis programs may offer options on how box plots are created. For example, the box plots in Figure 6 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).

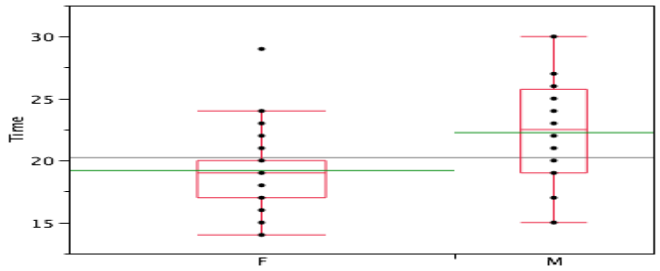


Figure 6. Box plots showing the individual scores and the means.

Each dot in Figure 6 represents a group of subjects with the same score (rounded to the nearest second). An alternative graphing technique is to jitter the points. 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 jittered. Figure 7 shows what jittering looks like.

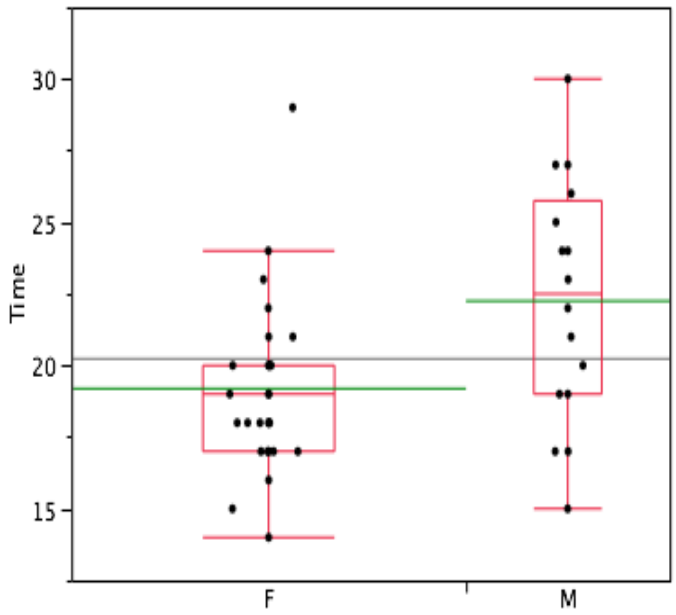


Figure 7. Box plots with the individual scores jittered.

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.

Bar Charts

by David M. Lane

Prerequisites

• Chapter 2: Graphing Qualitative Variables

Learning Objectives

1. Create and interpret bar charts
2. 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 1 shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

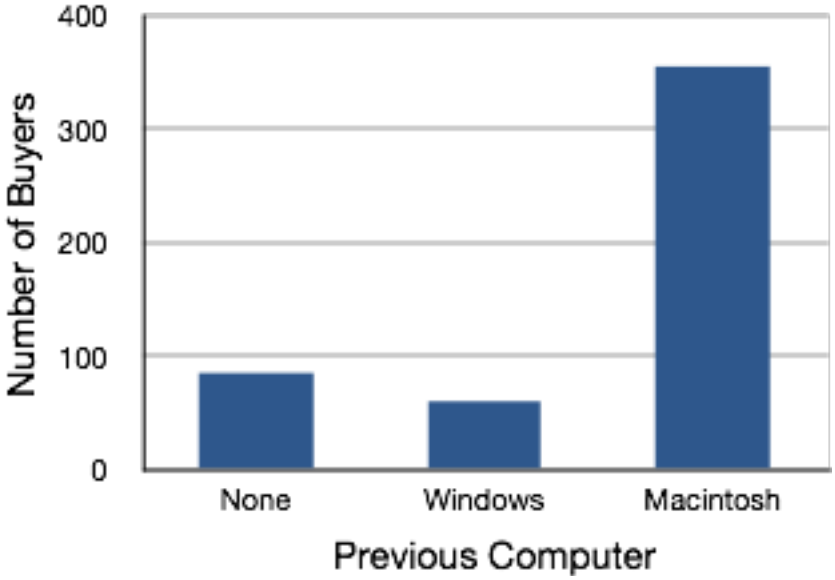


Figure 1. iMac buyers as a function of previous computer ownership.

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 shows the percent increases in the Dow Jones, Standard and Poor 500 (S & P), and Nasdaq stock indexes from May 24th 2000 to May 24th 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*.

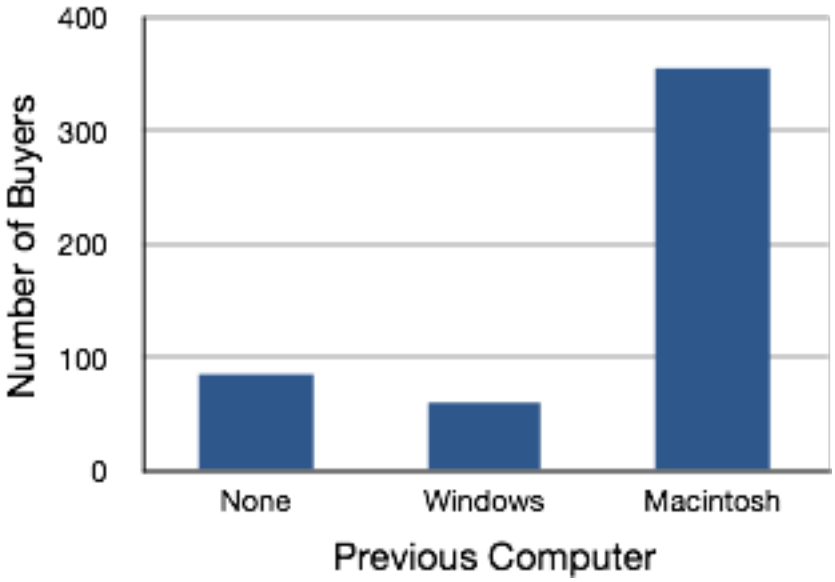


Figure 2. Percent increase in three stock indexes from May 24th 2000 to May 24th 2001.

Bar charts are particularly effective for showing change over time. Figure 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.

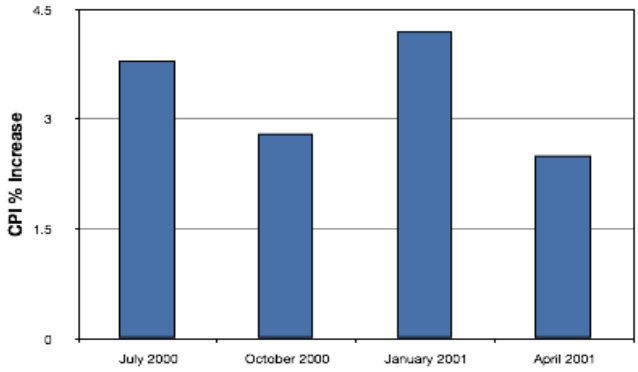


Figure 3. Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Bar charts are often used to compare the means of different experimental conditions. Figure 4 shows the mean time it took one of us (DL) to move the cursor

to either a small target or a large target. On average, more time was required for small targets than for large ones.

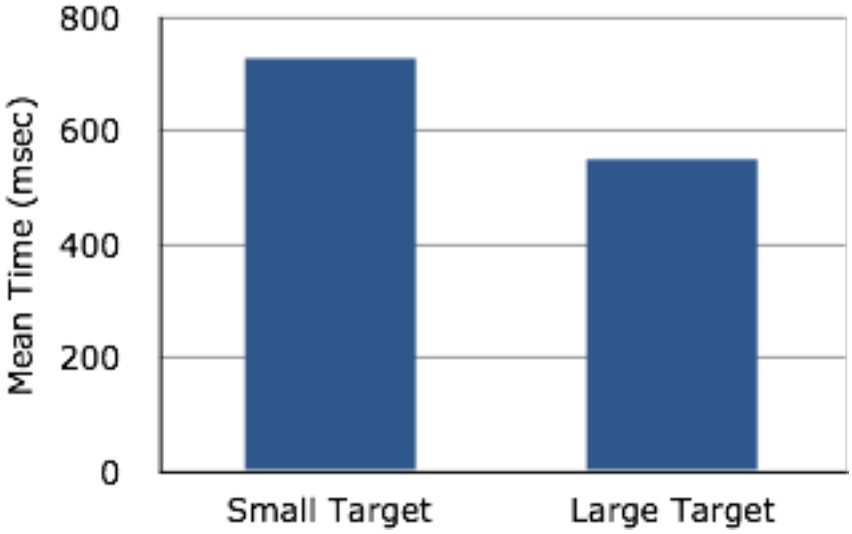


Figure 4. Bar chart showing the means for the two conditions.

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 cursor-movement data is shown in Figure 5. You can see that Figure 5 reveals more about the distribution of movement times than does Figure 4.

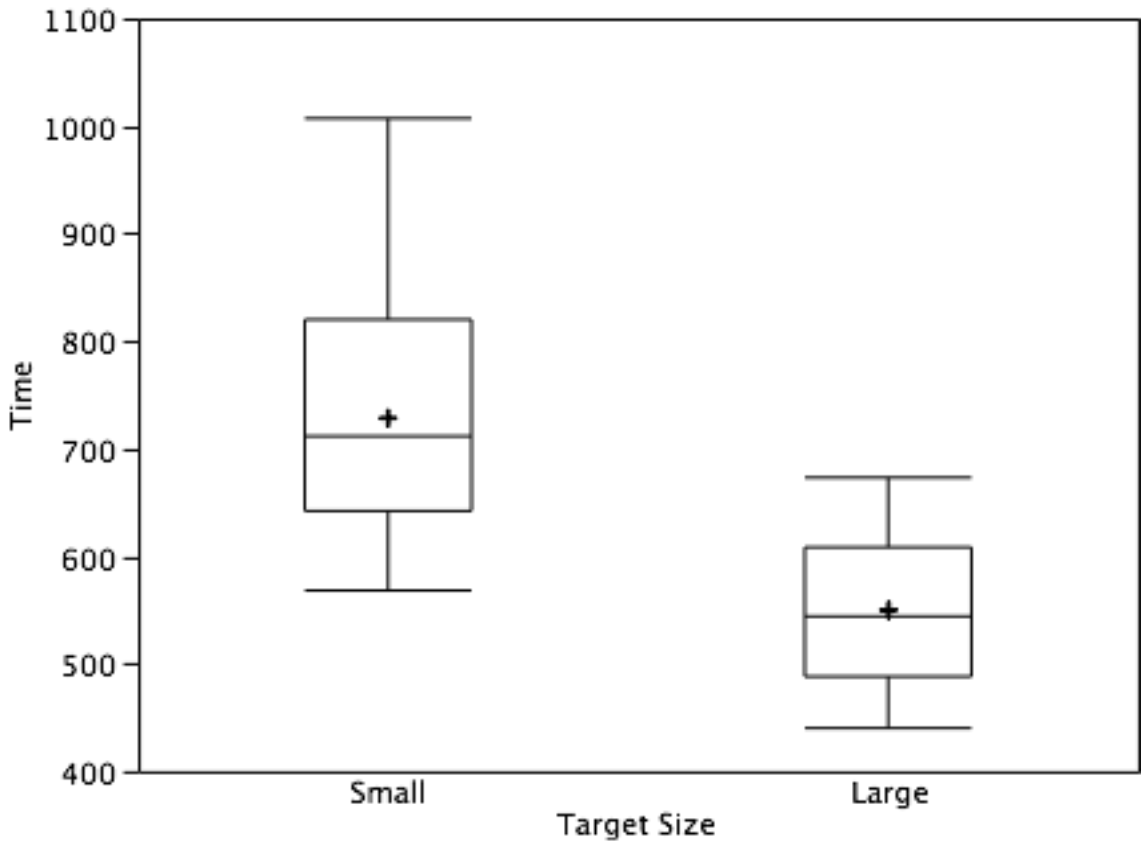


Figure 5. Box plots of times to move the cursor to the small and large targets.

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.

Line Graphs

by David M. Lane

Prerequisites

- Chapter 2: Bar Charts

Learning Objectives

1. Create and interpret line graphs
2. 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 1 was presented in the section on bar charts and shows changes in the Consumer Price Index (CPI) over time.

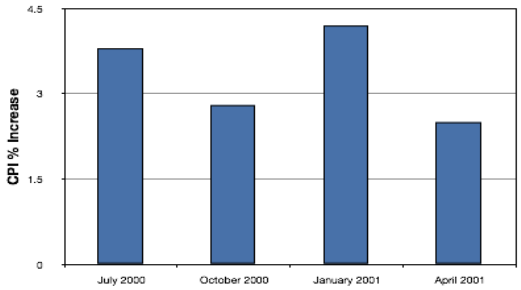


Figure 1. A bar chart of the percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

A line graph of these same data is shown in Figure 2. Although the figures are similar, the line graph emphasizes the change from period to period.

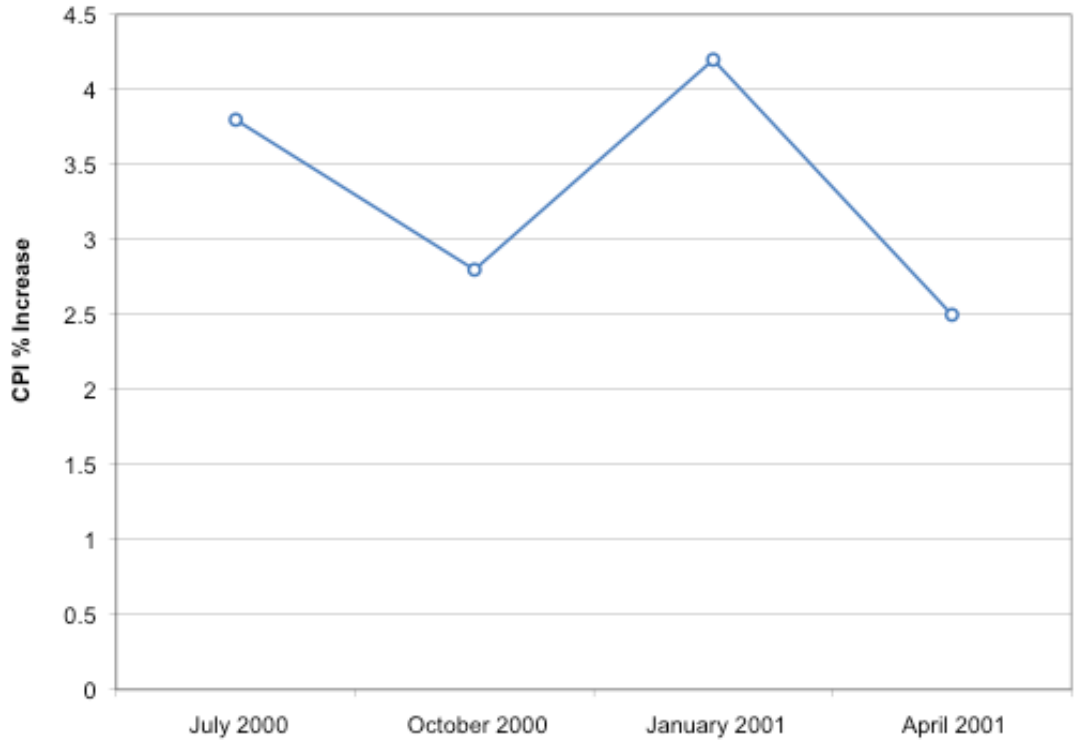


Figure 2. A line graph of the percent change in the CPI over time. Each point represents percent increase for the three months ending at the date indicated.

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 3, for example, shows percent increases and decreases in five components of the 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.

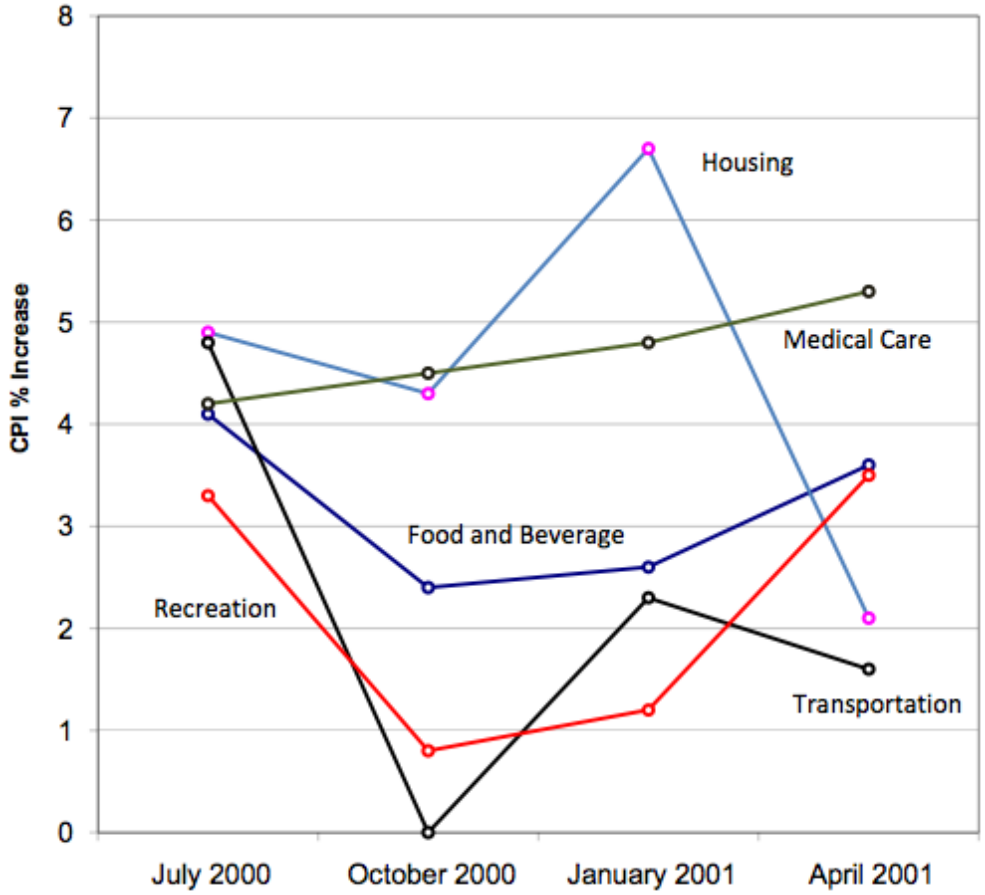


Figure 3. A line graph of the percent change in five components of the CPI over time.

Let us stress that it is misleading to use a line graph when the X-axis contains merely qualitative variables. Figure 4 inappropriately shows a line graph of the card game data from Yahoo, discussed in the section on qualitative variables. The defect in Figure 4 is that it gives the false impression that the games are naturally ordered in a numerical way.

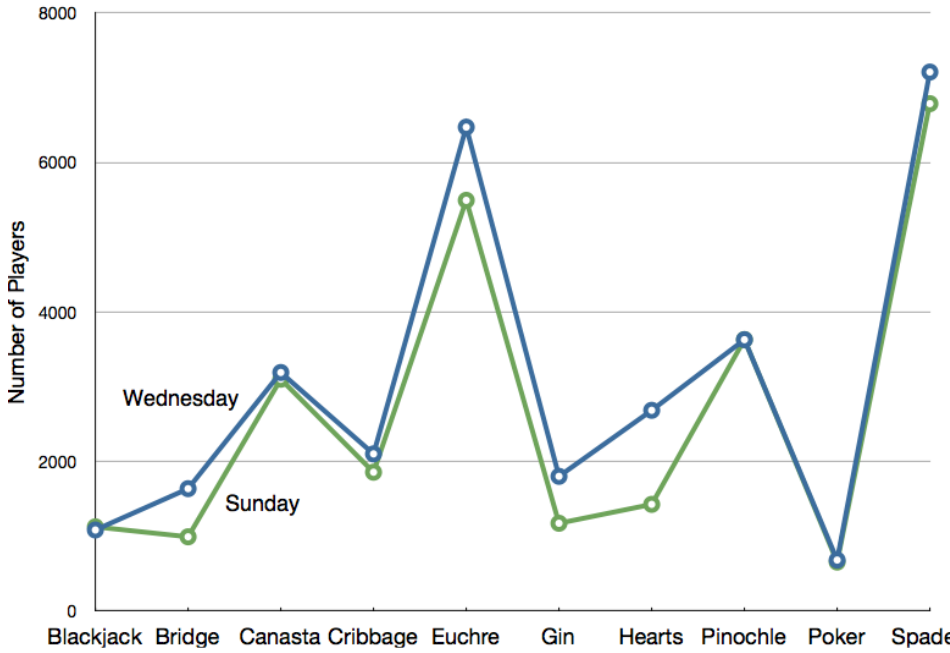


Figure 4. A line graph, inappropriately used, depicting the number of people playing different card games on Wednesday and Sunday.

Dot Plots

by David M. Lane

Prerequisites

- Chapter 2: Bar Charts

Learning Objectives

1. Create and interpret dot plots
2. 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 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.

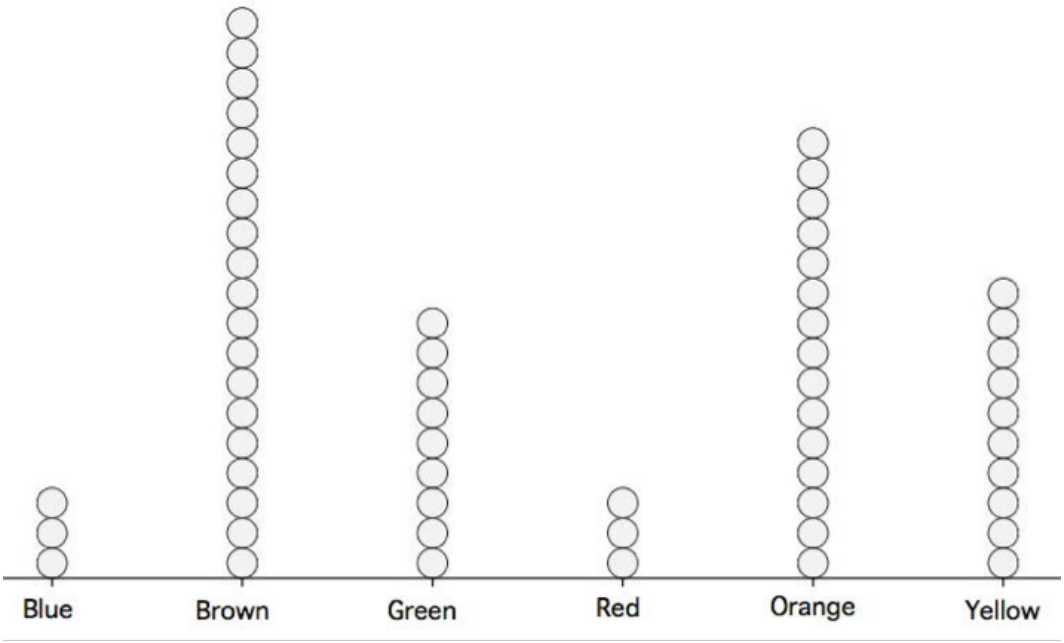


Figure 1. A dot plot showing the number of M & M's of various colors in a bag of M & M's.

The dot plot in Figure 2 shows the number of people playing various card games on the Yahoo website on a Wednesday. Unlike Figure 1, the location rather than the number of dots represents the frequency.

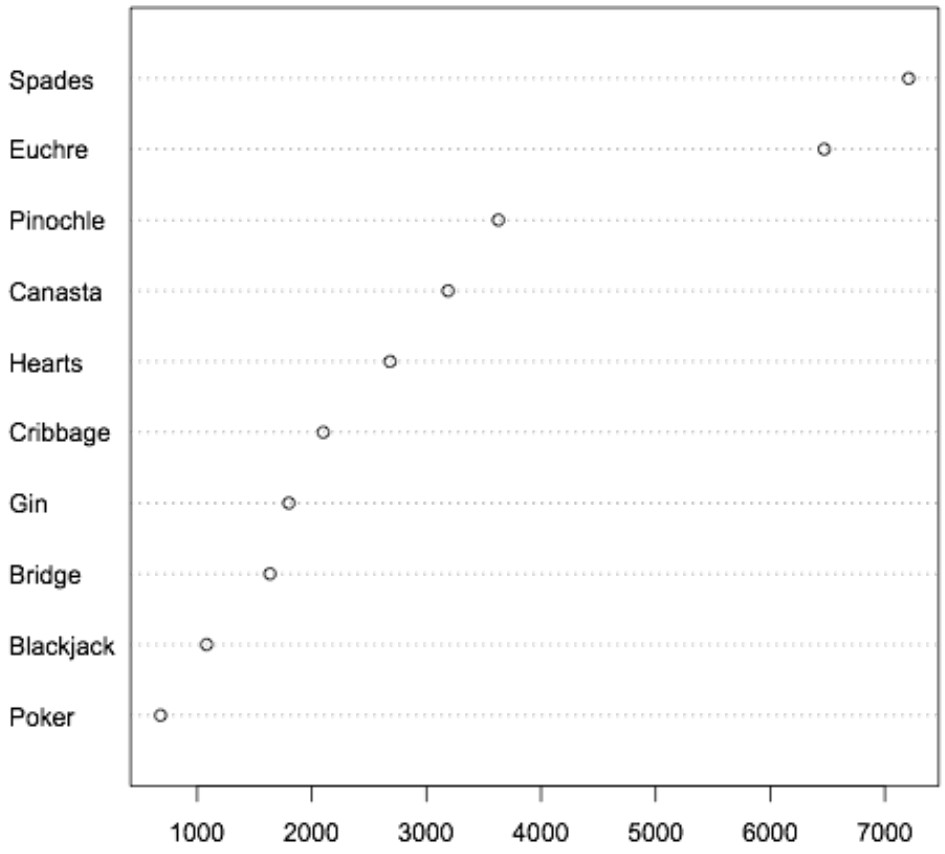


Figure 2. A dot plot showing the number of people playing various card games on a Wednesday.

The dot plot in Figure 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.

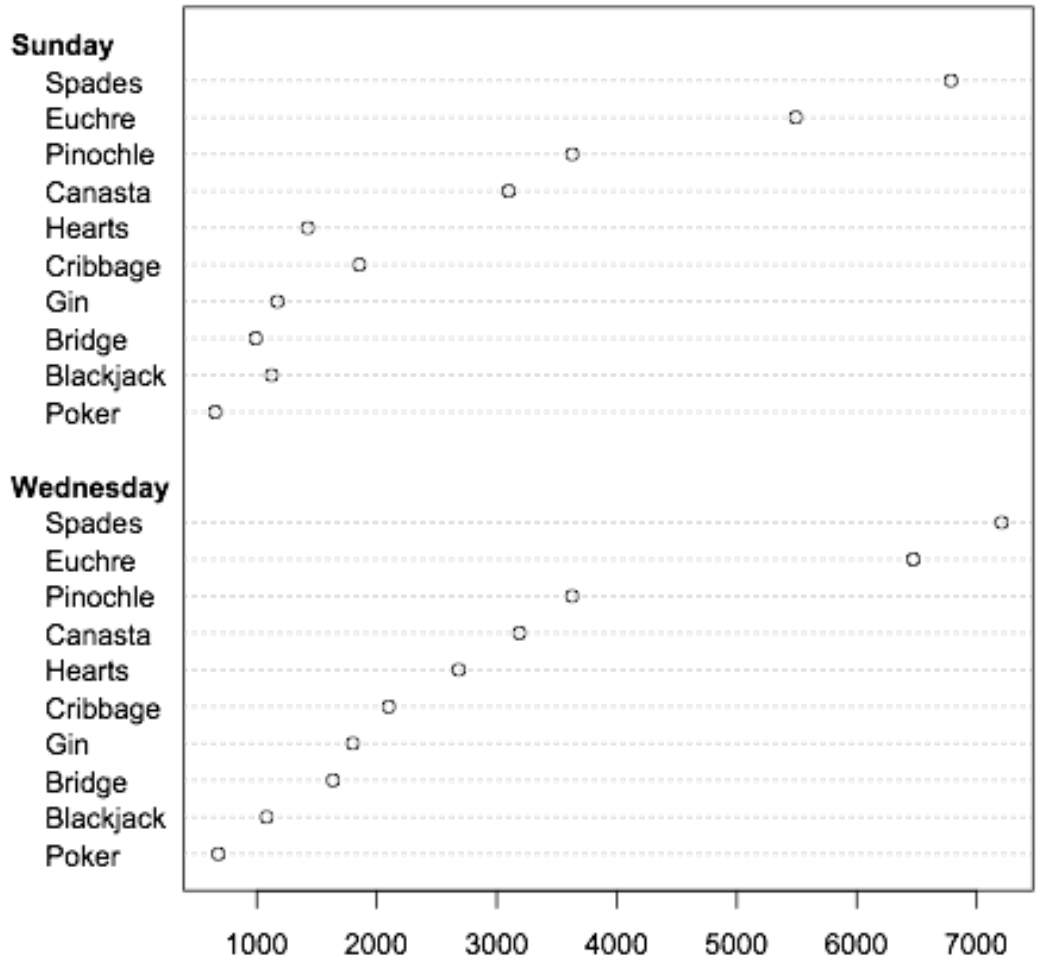


Figure 3. A dot plot showing the number of people playing various card games on a Sunday and on a Wednesday.

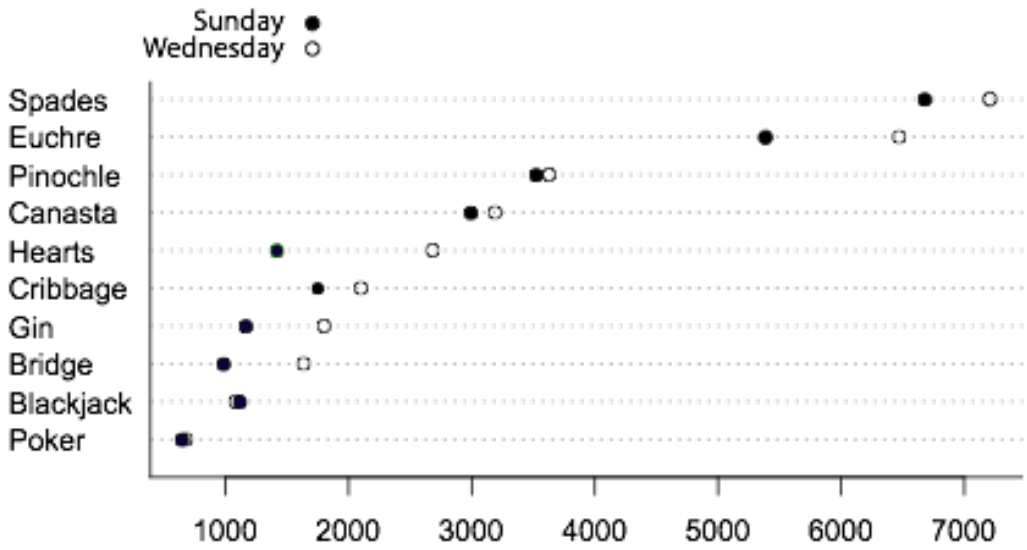


Figure 4. An alternate way of showing the number of people playing various card games on a Sunday and on a Wednesday.

The dot plot in Figure 4 makes it easy to compare the days of the week for specific games while still portraying differences among games.

3. Summarizing Distributions

- Central Tendency
  1. What is Central Tendency
  2. Measures of Central Tendency
  3. Median and Mean
  4. Additional Measures
  5. Comparing measures
- Variability
  1. Measures of Variability
- Shape
  1. Effects of Transformations
  2. Variance Sum Law I

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.

What is Central Tendency?

by David M. Lane and Heidi Ziemer

Prerequisites
- Chapter 1: Distributions
- Chapter 2: Stem and Leaf Displays

Learning Objectives
1. Identify situations in which knowing the center of a distribution would be valuable
2. Give three different ways the center of a distribution can be defined
3. 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 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 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 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.

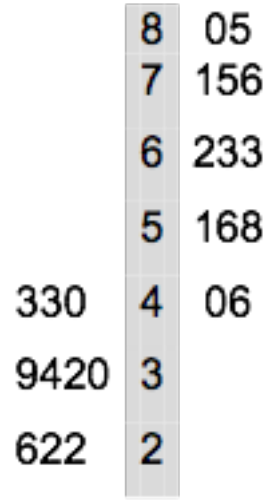


Figure 1. Back-to-back stem and leaf display. The left side shows the memory scores of the non-players. The right side shows the scores of the tournament players.

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.

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 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.

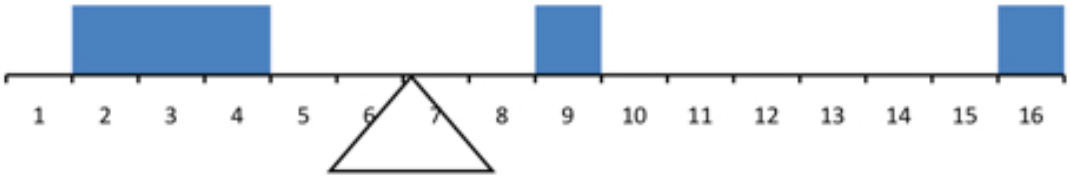


Figure 2. A balance scale.

For another example, consider the distribution shown in Figure 3. It is balanced by placing the fulcrum in the geometric middle.

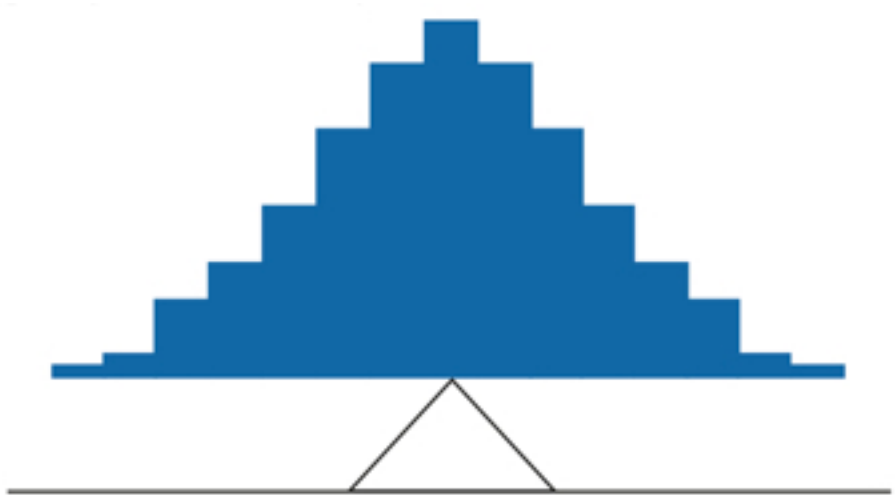


Figure 3. A distribution balanced on the tip of a triangle.

Figure 4 illustrates that the same distribution can't be balanced by placing the fulcrum to the left of center.

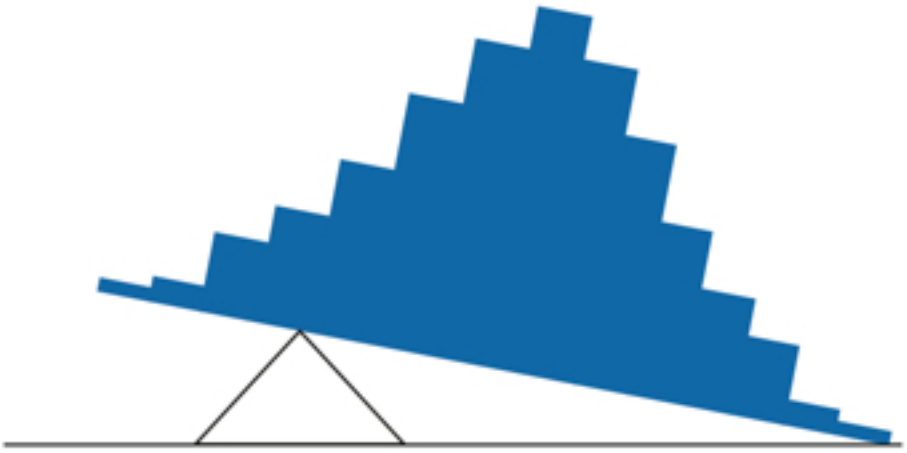


Figure 4. The distribution is not balanced.

Figure 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). Placing the fulcrum at the "half way" point would cause it to tip towards the left.

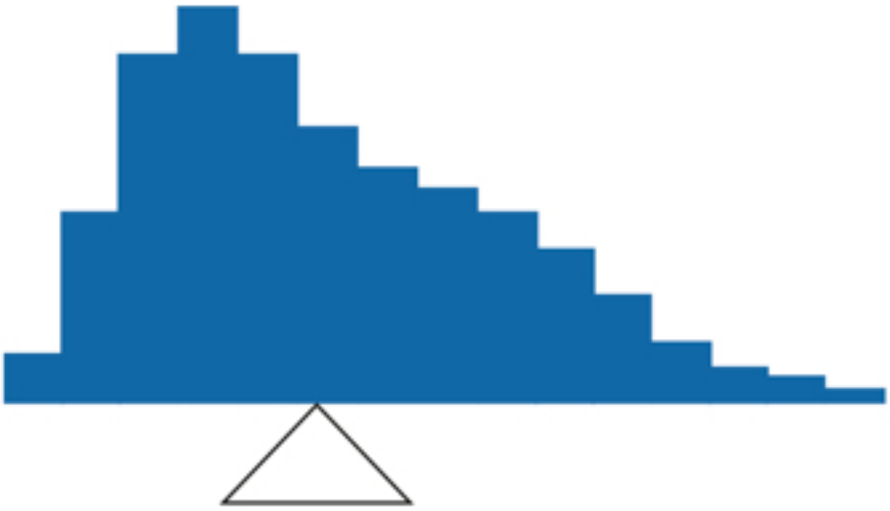


Figure 5. An asymmetric distribution balanced on the tip of a triangle.

The balance point defines one sense of a distribution's center.

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 2 shows the sum of the absolute deviations of these numbers from the number 10.

Table 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.

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 shows the sum of the squared deviations of these numbers from the number 10.

Table 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.

Measures of Central Tendency

by David M. Lane

Prerequisites
- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: Central Tendency

Learning Objectives
1. Compute mean
2. Compute median
3. 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.

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:

where ΣX 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:

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

Table 1. Number of touchdown passes.

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

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 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:

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 1, 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 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 2. Grouped frequency distribution

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

Median and Mean

by David M. Lane

Prerequisites

- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency

Learning Objectives

1. State when the mean and median are the same
2. State whether it is the mean or median that minimizes the mean absolute deviation
3. State whether it is the mean or median that minimizes the mean squared deviation
4. 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, and (3) the value whose squared difference from all the other values is minimized. 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 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 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 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.

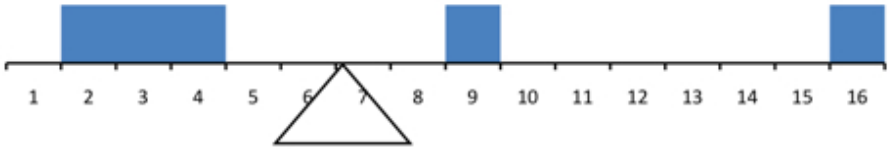


Figure 1. The distribution balances at the mean of 6.8 and not at the median of 4.0.

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.

Additional Measures of Central Tendency

by David M. Lane

Prerequisites

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency
- Chapter 3: Mean and Median

Learning Objectives

1. Compute the trimean
2. Compute the geometric mean directly
3. Compute the geometric mean using logs
4. Use the geometric to compute annual portfolio returns
5. 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:

$$Trimean = \frac{(P25 + 2P50 + P75)}{4}$$

Consider the data in Table 2. The 25th percentile is 15, the 50th is 20 and the 75th percentile is 23.

Table 1. Number of touchdown passes.

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

Table 2. Percentiles.

Percentile	Value
25	15
50	20
75	23

The trimean is therefore :

$$\frac{(15 + 2 \times 20 + 23)}{4} = \frac{78}{4} = 19.5$$

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 x 10 x 100 = 1,000. 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

$$\left(\prod X\right)^{\frac{1}{N}}$$

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/Nth power. Raising a value to the 1/Nth 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 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. Logarithms.

X	Log10(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 4 shows the value after each of the five years.

Table 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 5 shows how a portfolio gaining 5% a year would end up with the same value ($1,276) as shown in Table 4.

Table 5. Portfolio Returns

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

Comparing Measures of Central Tendency

by David M. Lane

Prerequisites

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency
- Chapter 3: Mean and Median

Learning Objectives

1. Understand how the difference between the mean and median is affected by skew
2. State how the measures differ in symmetric distributions
3. 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 1 shows the distribution of 642 scores on an introductory psychology test. Notice this distribution has a slight positive skew.

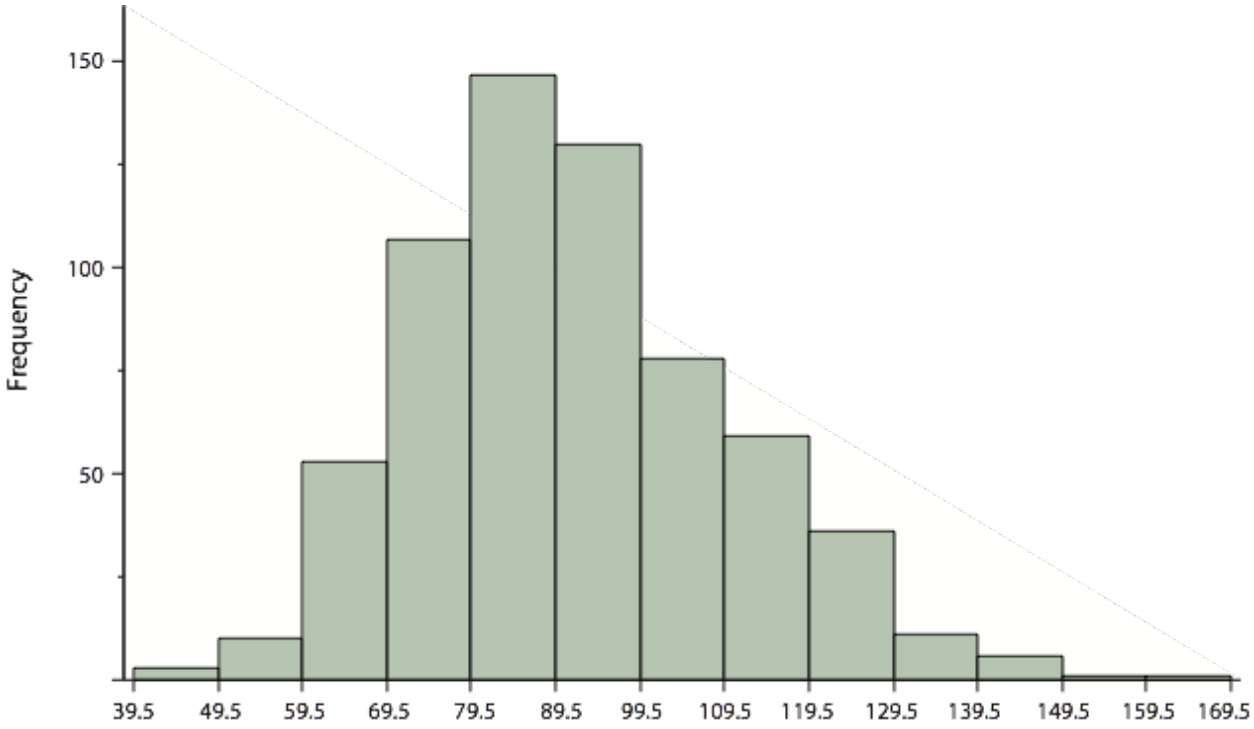


Figure 1. A distribution with a positive skew.

Measures of central tendency are shown in Table 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 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 2 has a much more pronounced skew than the distribution in Figure 1.

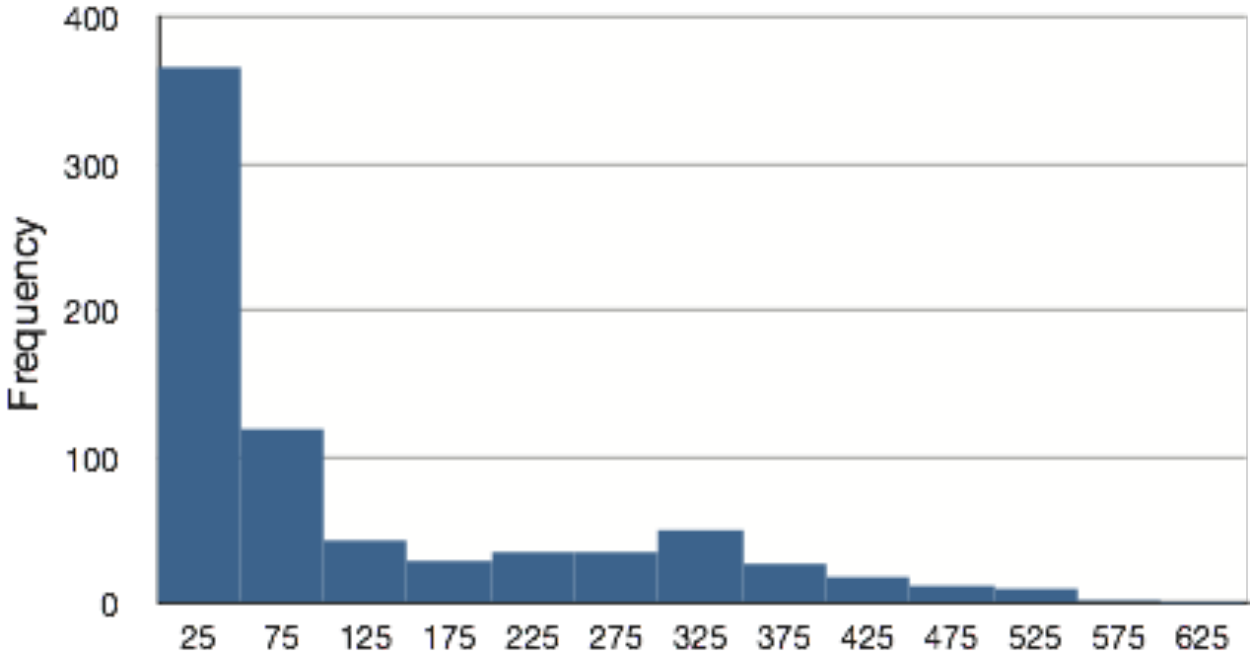


Figure 2. A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

Table 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 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

Measures of Variability

by David M. Lane

Prerequisites

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: Measures of Central Tendency

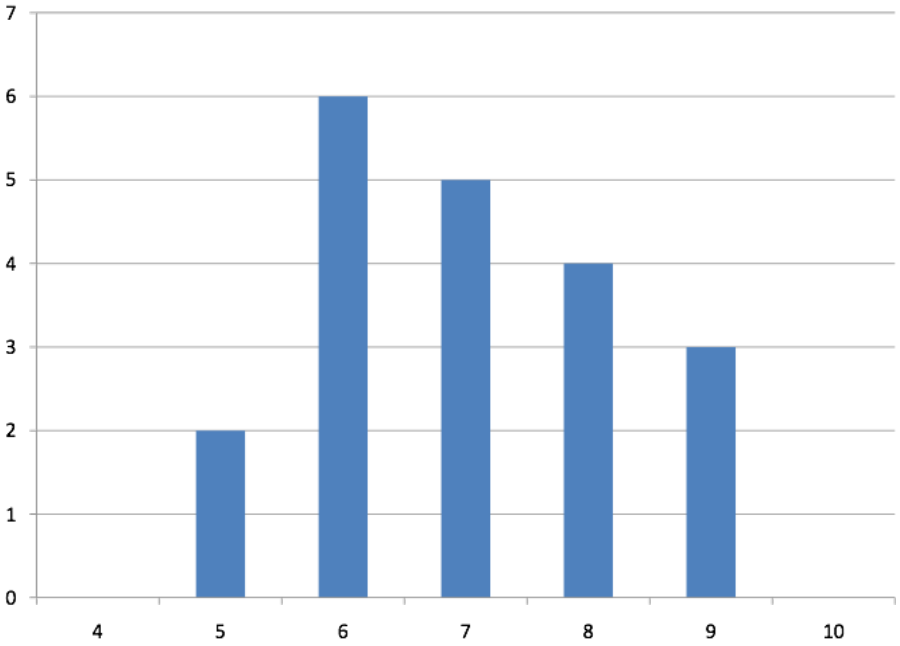
Learning Objectives

1. Determine the relative variability of two distributions
2. Compute the range
3. Compute the inter-quartile range
4. Compute the variance in the population
5. Estimate the variance from a sample
6. 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 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.

Quiz 1



Quiz 2

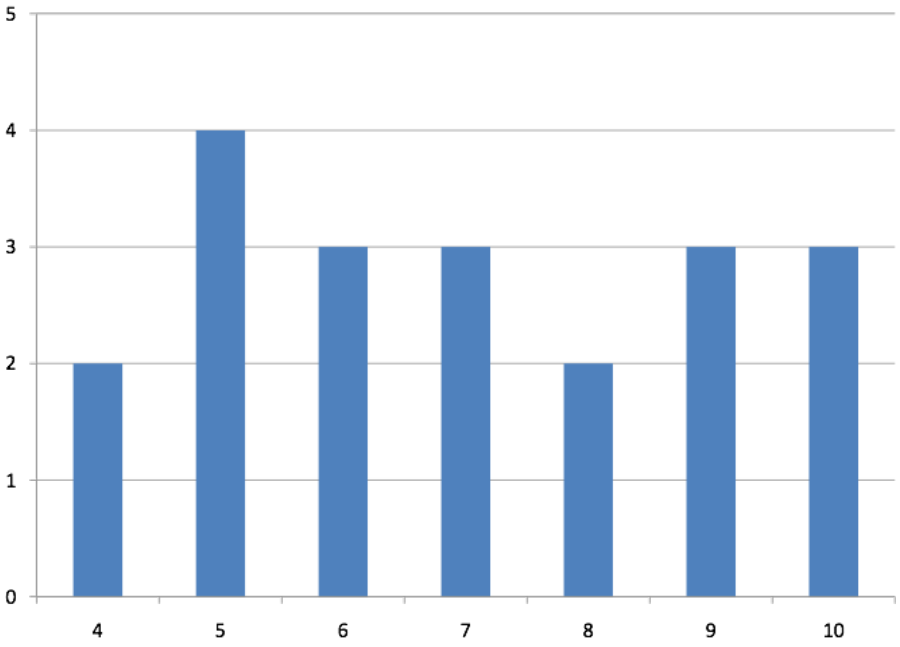


Figure 1. Bar charts of two quizzes.

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: 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 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 = 75th percentile - 25th percentile
```

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 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 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 Quiz 2 show that its variance is 6.7. The formula for the variance is:

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

where σ^2 is the variance, μ is the mean, and N is the number of numbers. For 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:

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

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.

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:

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

$$s^2 = \frac{(1-3)^2+(2-3)^2+(4-3)^2+(5-3)^2}{4-1} = \frac{4+1+1+4}{3} = \frac{10}{3} = 3.333$$

There are alternate formulas that can be easier to use if you are doing your calculations with a hand calculator:

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

and

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

For this example,

$$\left(\sum X\right)^2 = \frac{(1+2+4+5)^2}{4} = \frac{144}{4} = 36$$

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

$$s^2 = \frac{(46-36)}{3} = 3.333$$

as with the other formula.

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 7) 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 x 10 = 30 and 50 + 2 x 10 = 70. The symbol for the population standard deviation is σ; the symbol for an estimate computed in a sample is s. Figure 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 45 and 55; for the blue distribution, 68% is between 40 and 60.

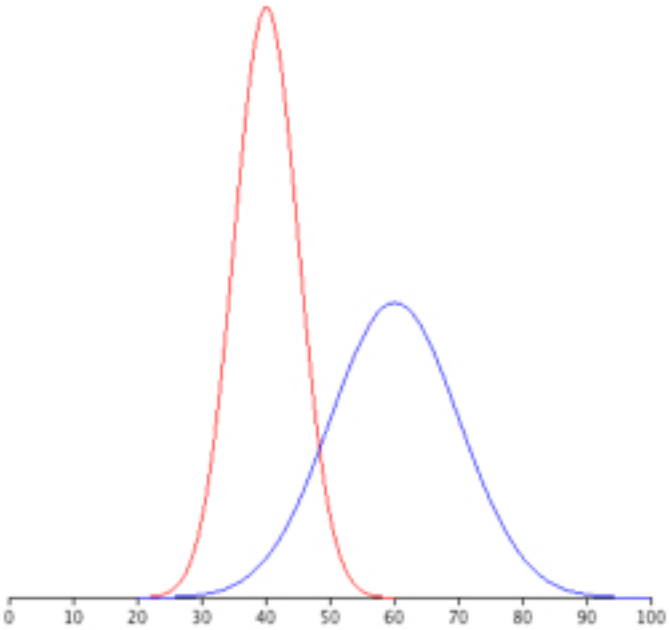


Figure 2. Normal distributions with standard deviations of 5 and 10.

Shapes of Distributions

by David M. Lane

Prerequisites

- Chapter 1: Distributions
- Chapter 3: Measures of Central Tendency
- Chapter 3: Variability

Learning Objectives

1. Compute skew using two different formulas
2. 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 1 shows a distribution with a very large positive skew. Recall that distributions with positive skew have tails that extend to the right.

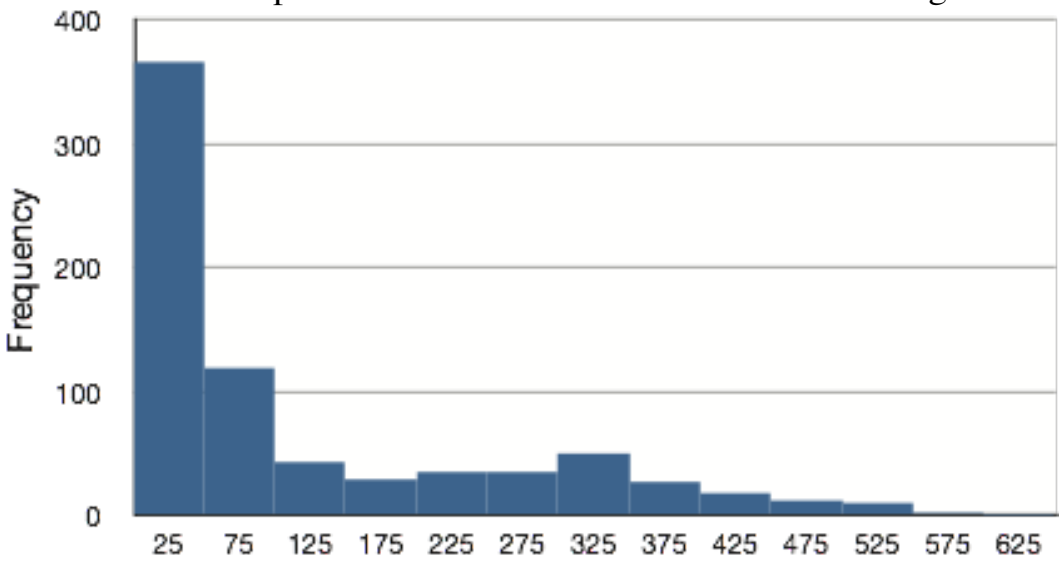


Figure 1. A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

Distributions with positive skew normally have larger means than medians. The mean and median of the baseball salaries shown in Figure 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 lead the statistician Pearson to propose the following simple and convenient numerical index of skew:

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

The standard deviation of the baseball salaries is 1,390,922. Therefore, Pearson's measure of skew for this distribution is 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.

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

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.

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

Effects of Linear Transformations

by David M. Lane

Prerequisites

• Chapter 1: Linear Transformations

Learning Objectives

1. Define a linear transformation
2. Compute the mean of a transformed variable
3. 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 1 shows the temperatures of 5 cities.

Table 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

$$C = 0.55556F - 17.7778$$

which means we multiply each temperature Fahrenheit by 0.556 and then subtract 17.7778. 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) - 17.7778 = 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 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 0.556^2 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

$$Y = bX + A$$

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.

Variance Sum Law I

by David M. Lane

Prerequisites
• Chapter 3: Variance

Learning Objectives
1. Compute the variance of the sum of two uncorrelated variables
2. Compute the variance of the difference between two uncorrelated variables

As you will see in later sections, there are many occasions in which it is important to know the variance of the sum of two variables. Consider the following situation: (a) you have two populations, (b) you sample one number from each population, and (c) you add the two numbers together. The question is, "What is the variance of this sum?" For example, suppose the two populations are the populations of 8-year old males and 8-year-old females in Houston, Texas, and that the variable of interest is memory span. You repeat the following steps thousands of times: (1) sample one male and one female, (2) measure the memory span of each, and (3) sum the two memory spans. After you have done this thousands of times, you compute the variance of the sum. It turns out that the variance of this sum can be computed according to the following formula:

$$\sigma^2_{sum} = \sigma^2_M + \sigma^2_F$$

where the first term is the variance of the sum, the second term is the variance of the males and the third term is the variance of the females. Therefore, if the variances on the memory span test for the males and females respectively were 0.9 and 0.8, respectively, then the variance of the sum would be 1.7.

The formula for the variance of the difference between the two variables (memory span in this example) is shown below. Notice that the expression for the difference is the same as the formula for the sum.

$$\sigma^2_{difference} = \sigma^2_M + \sigma^2_F$$

More generally, the variance sum law can be written as follows:

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$$

which is read: “The variance of X plus or minus Y is equal to the variance of X plus the variance of Y.”

These formulas for the sum and difference of variables given above only apply when the variables are independent.

In this example, we have thousands of randomly-paired scores. Since the scores are paired randomly, there is no relationship between the memory span of one member of the pair and the memory span of the other. Therefore the two scores are independent. Contrast this situation with one in which thousands of people are sampled and two measures (such as verbal and quantitative SAT) are taken from each. In this case, there would be a relationship between the two variables since higher scores on the verbal SAT are associated with higher scores on the quantitative SAT (although there are many examples of people who score high on one test and low on the other). Thus the two variables are not independent and the variance of the total SAT score would not be the sum of the variances of the verbal SAT and the quantitative SAT. The general form of the variance sum law is presented in a section in the chapter on correlation.

4. Describing Bivariate Data

- Introduction to Bivariate Data
- Values of the Pearson Correlation
- Properties of Pearson's r
- Computing Pearson's r
- Variance Sum Law II

A dataset with two variables contains what is called bivariate data. This chapter discusses ways to describe the relationship between two variables. For example, you may wish to describe the relationship between the heights and weights of people to determine the extent to which taller people weigh more.

The introductory section gives more examples of bivariate relationships and presents the most common way of portraying these relationships graphically. The next five sections discuss Pearson's correlation, the most common index of the relationship between two variables. The final section, “Variance Sum Law II,” makes use of Pearson's correlation to generalize this law to bivariate data.

Introduction to Bivariate Data

by Rudy Guerra and David M. Lane

Prerequisites

- Chapter 1: Variables
- Chapter 1: Distributions
- Chapter 2: Histograms
- Chapter 3: Measures of Central Tendency
- Chapter 3: Variability
- Chapter 3: Shapes of Distributions

Learning Objectives

1. Define “bivariate data”
2. Define “scatter plot”
3. Distinguish between a linear and a nonlinear relationship
4. 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 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 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 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 1) and by a mean and standard deviation (See Table 2).

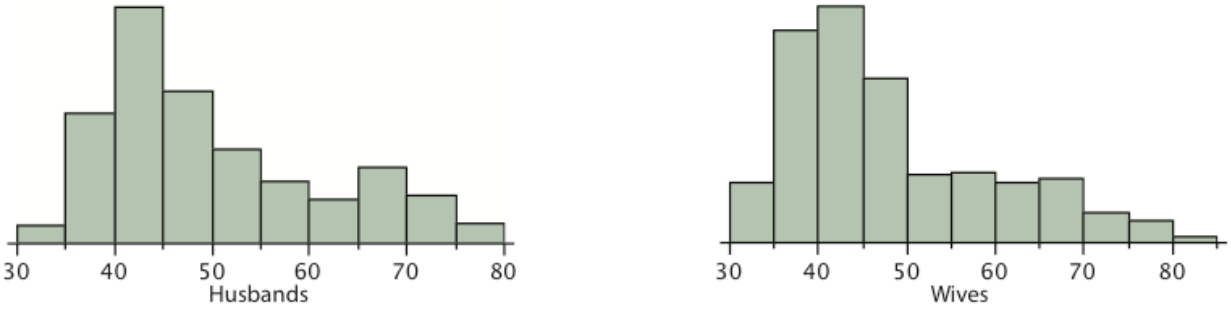


Figure 1. Histograms of spousal ages.

Table 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 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 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.

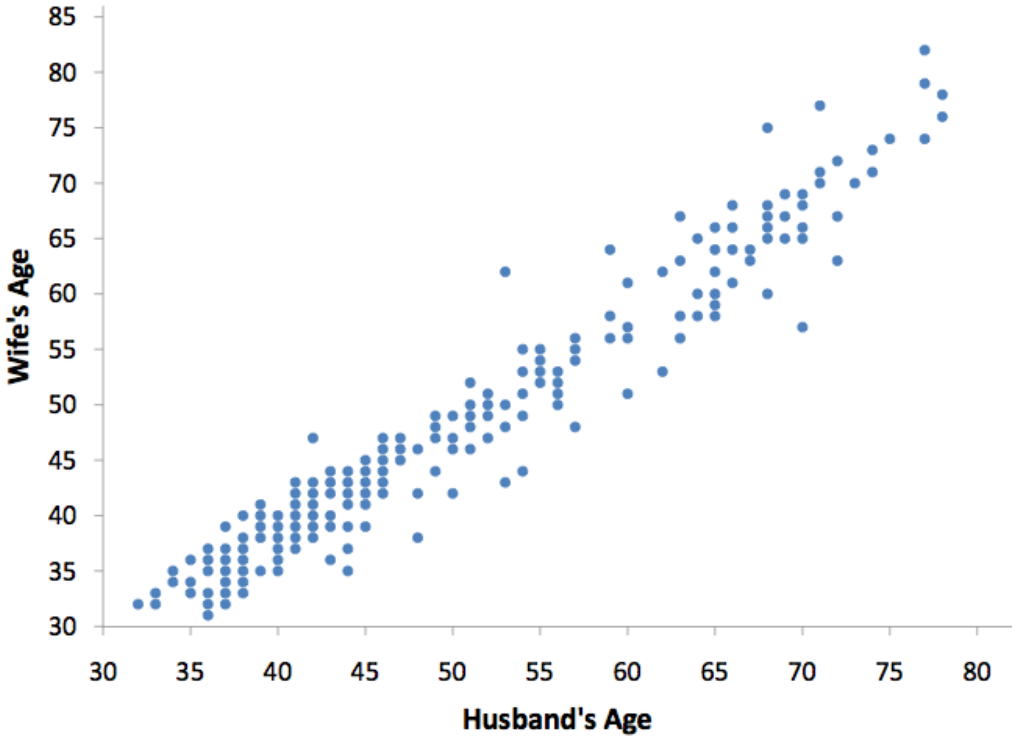


Figure 2. Scatter plot showing wife's age as a function of husband's age.

There are two important characteristics of the data revealed by Figure 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 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.

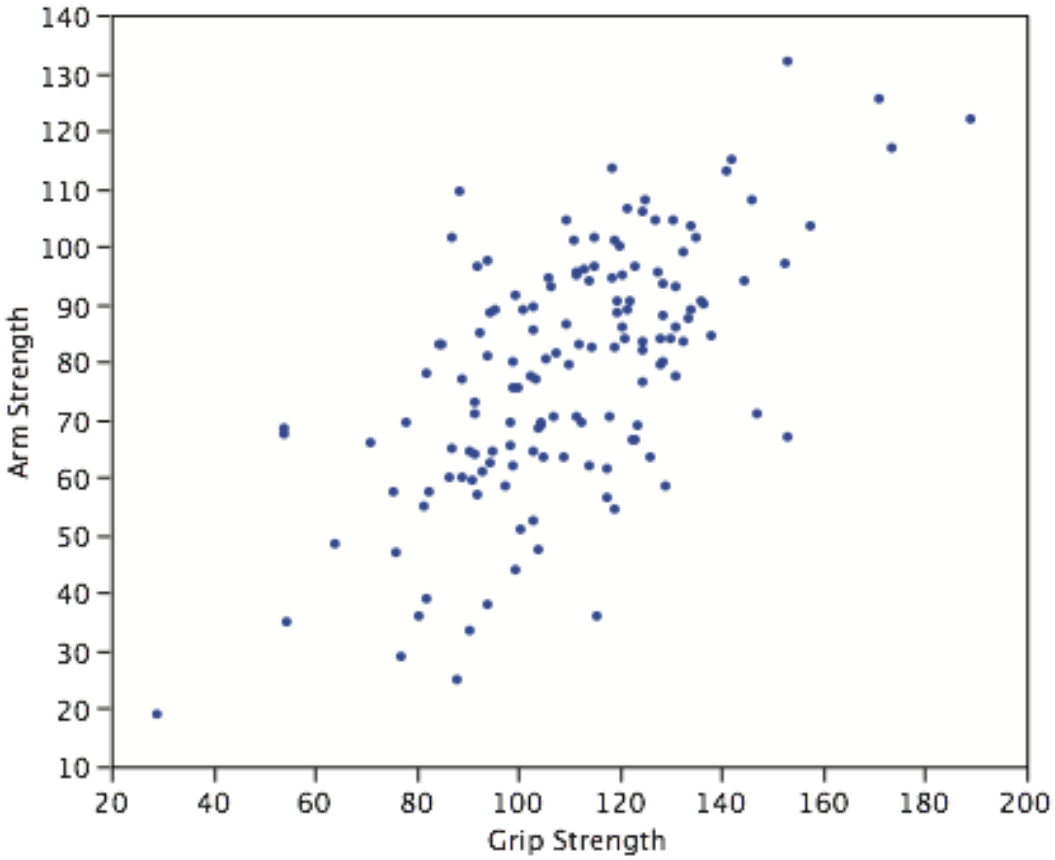


Figure 3. Scatter plot of Grip Strength and Arm Strength.

Not all scatter plots show linear relationships. Figure 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 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.

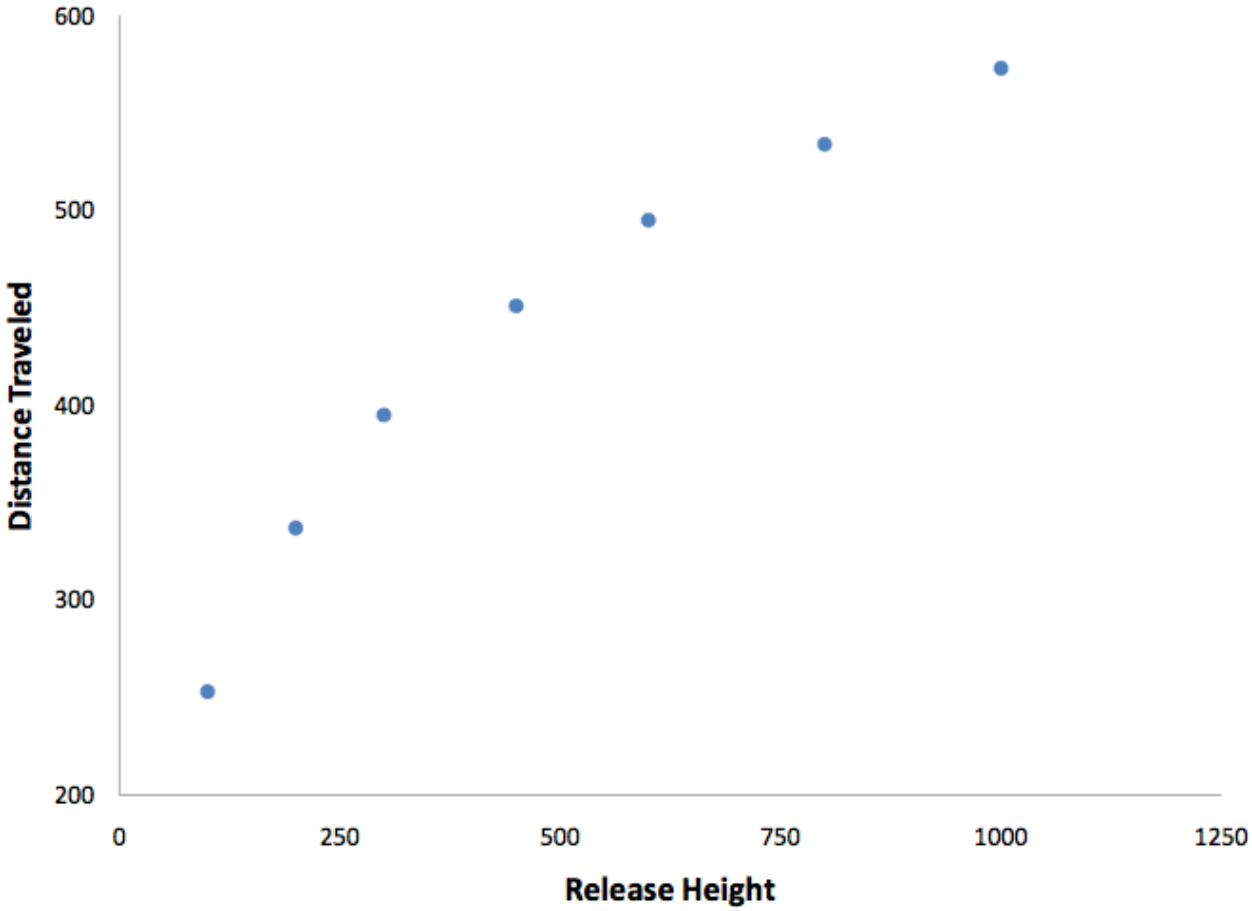


Figure 4. Galileo's data showing a non-linear relationship.

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 next section.

Values of the Pearson Correlation

by David M. Lane

Prerequisites

• Chapter 4: Introduction to Bivariate Data

Learning Objectives

1. Describe what Pearson's correlation measures
2. Give the symbols for Pearson's correlation in the sample and in the population
3. State the possible range for Pearson's correlation
4. 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 1 shows a scatter plot for which r = 1.

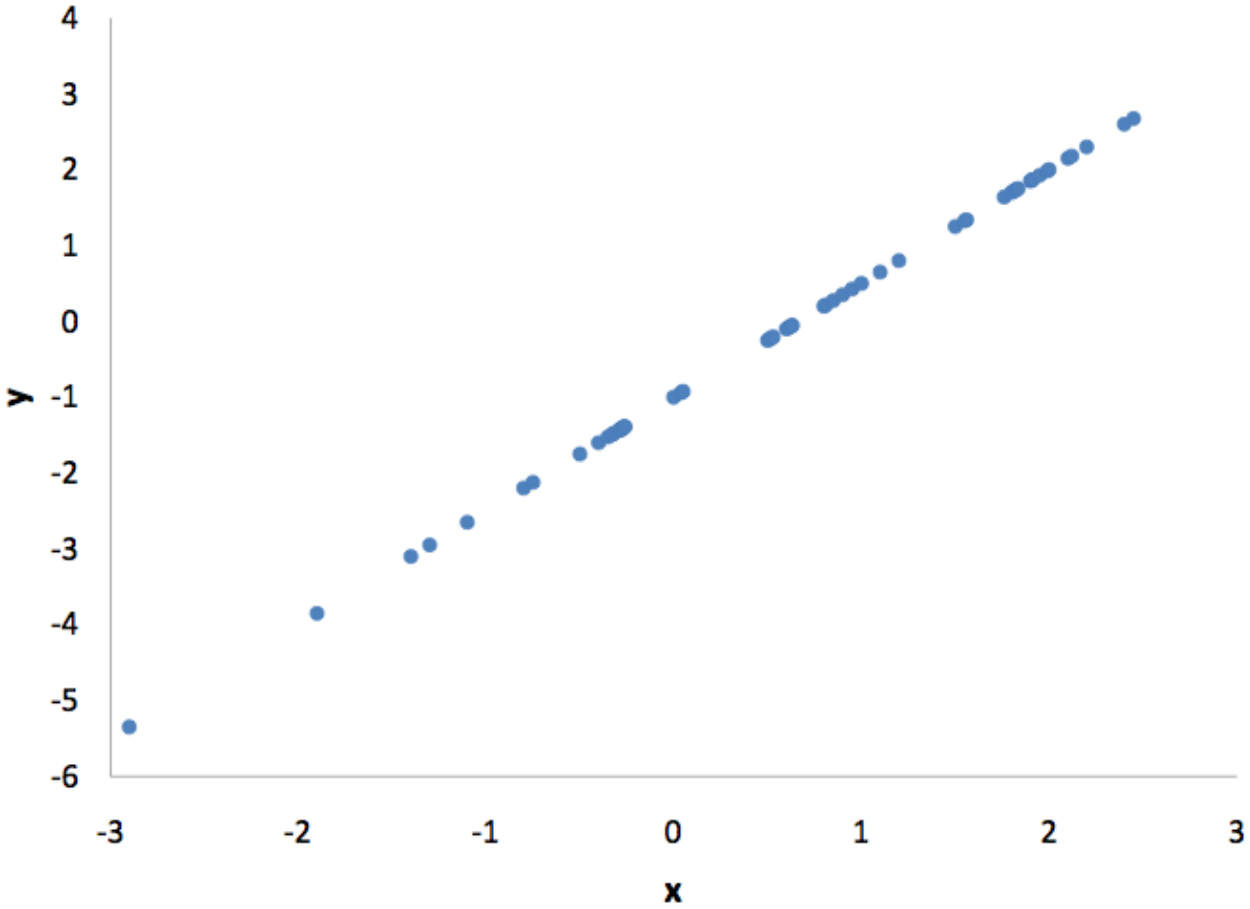


Figure 1. A perfect linear relationship, r = 1.

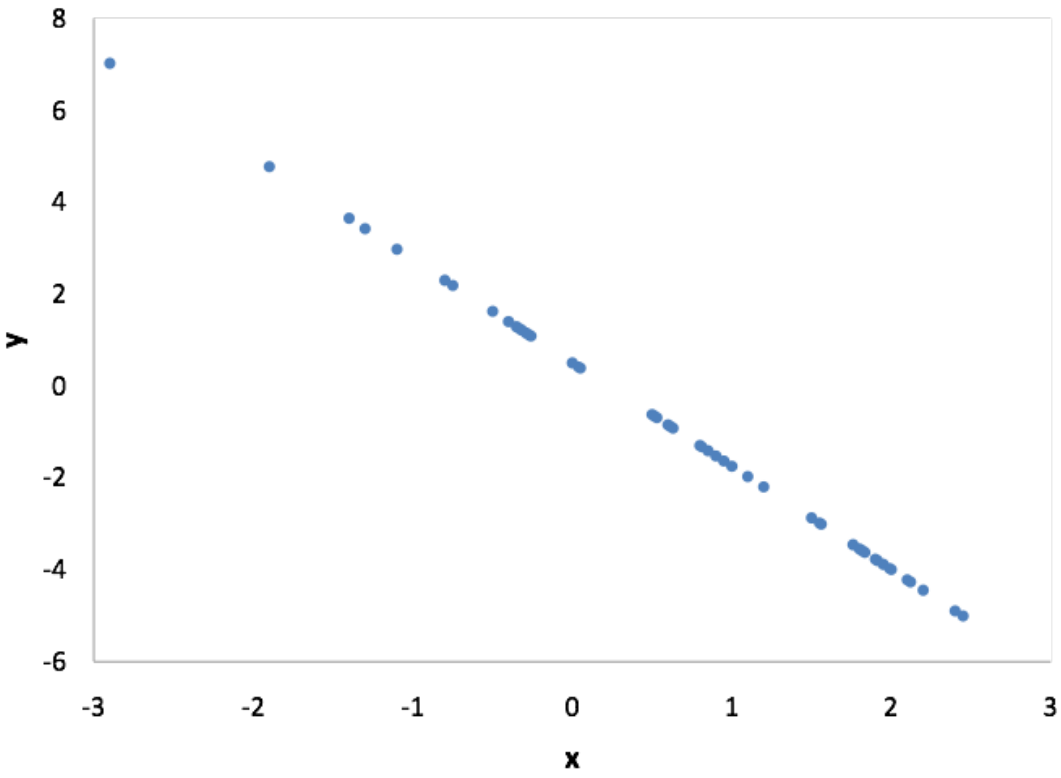


Figure 2. A perfect negative linear relationship, r = -1.

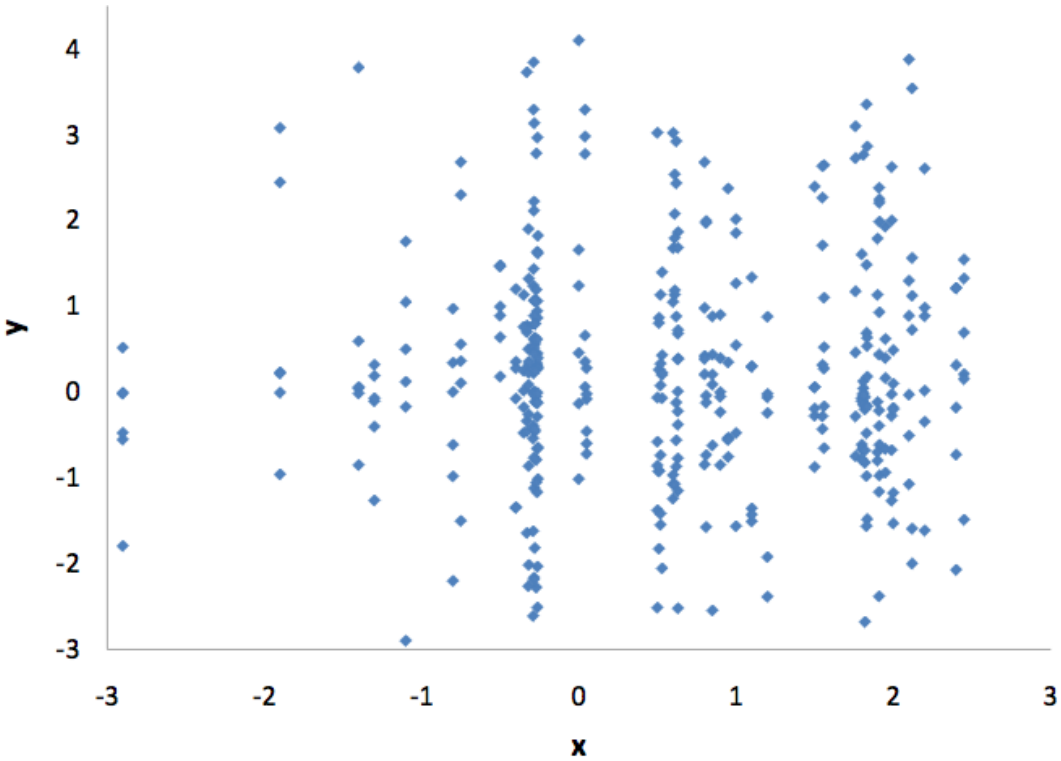


Figure 3. A scatter plot for which r = 0. Notice that there is no relationship between X and Y.

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 4 and described in the introductory section has an r of 0.97.

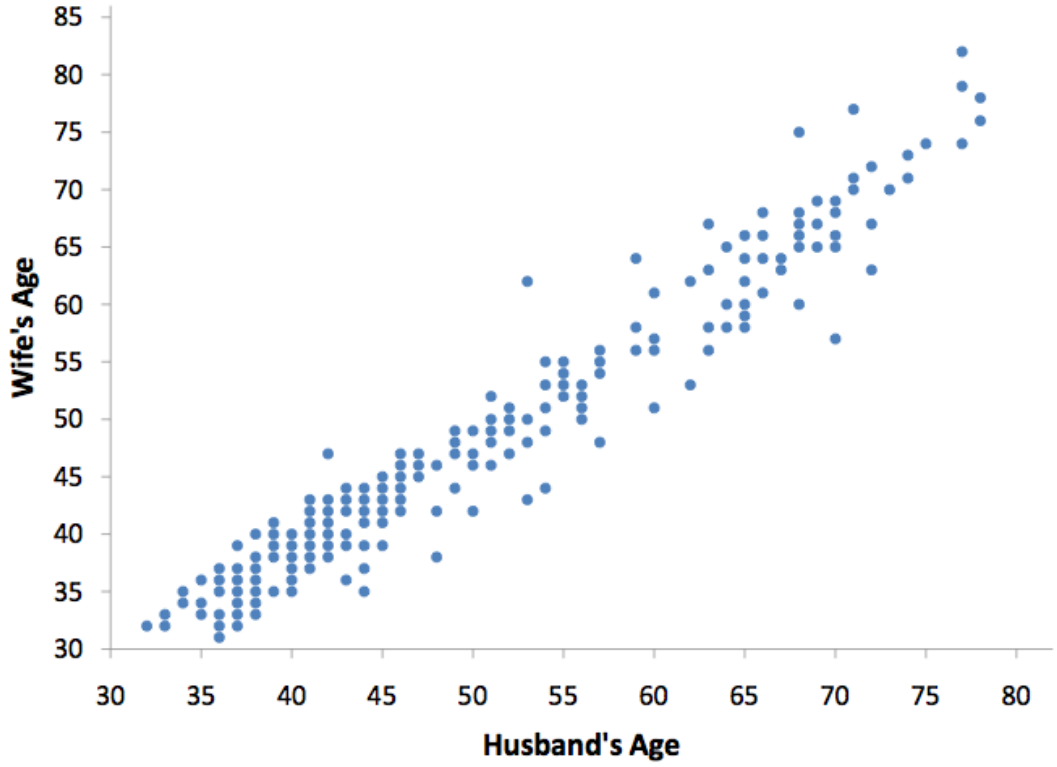


Figure 4. Scatter plot of spousal ages, r = 0.97.

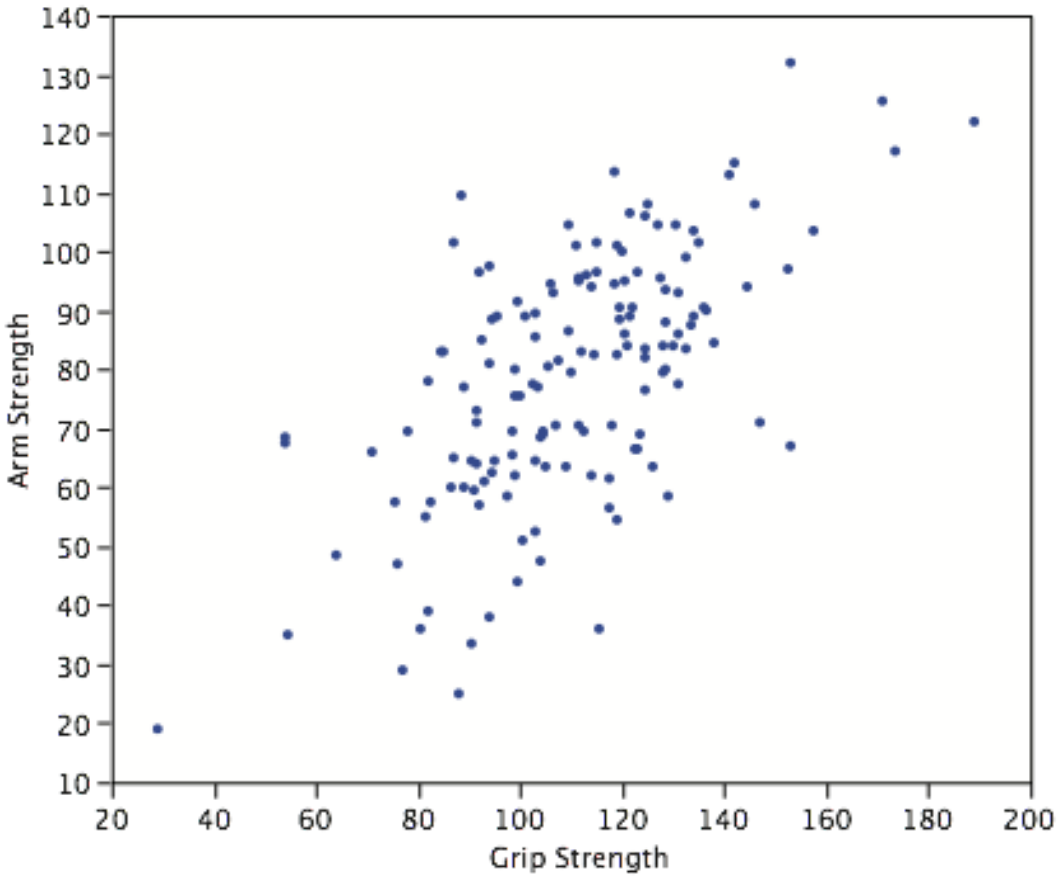


Figure 5. Scatter plot of Grip Strength and Arm Strength, r = 0.63.

The relationship between grip strength and arm strength depicted in Figure 5 (also described in the introductory section) is 0.63.

Properties of Pearson's r

by David M. Lane

Prerequisites

- Chapter 1: Linear Transformations
- Chapter 4: Introduction to Bivariate Data

Learning Objectives

1. State the range of values for Pearson's correlation
2. State the values that represent perfect linear relationships
3. State the relationship between the correlation of Y with X and the correlation of X with Y
4. 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.

Computing Pearson's r

by David M. Lane

Prerequisites

- Chapter 1: Summation Notation
- Chapter 4: Introduction to Bivariate Data

Learning Objectives

1. Define X and x
2. State why $\Sigma xy = 0$ when there is no relationship
3. 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 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 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 (Σxy) by the square root of the product of the sum of the x^2 column (Σx^2) and the sum of the y^2 column (Σy^2). The resulting formula is:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

and therefore

$$r = \frac{30}{\sqrt{(16)(60)}} = \frac{30}{\sqrt{960}} = \frac{30}{30.984} = 0.968$$

An alternative computational formula that avoids the step of computing deviation scores is:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{N}\right)}\sqrt{\left(\sum y^2 - \frac{(\sum y)^2}{N}\right)}}$$

Variance Sum Law II

by David M. Lane

Prerequisites
- Chapter 1: Variance Sum Law I
- Chapter 4: Values of Pearson's Correlation

Learning Objectives
1. State the variance sum law when X and Y are not assumed to be independent
2. Compute the variance of the sum of two variables if the variance of each and their correlation is known
3. 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:

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$$

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:

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho\sigma_X\sigma_Y$$

where ϱ is the correlation between X and Y in the population. For example, 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, then the variance of total SAT (verbal + quantitative) would be:

$$\sigma^2_{verbal+quant} = 10{,}000 + 11{,}000 + (2)(0.5)\sqrt{10{,}000}\sqrt{11{,}000}$$

which is equal to 31,488. The variance of the difference is:

$$\sigma^2_{verbal-quant} = 10{,}000 + 11{,}000 - (2)(0.5)\sqrt{10{,}000}\sqrt{11{,}000}$$

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:

$$s^2_{X \pm Y} = s^2_X + s^2_Y \pm 2r s_X s_y$$