UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Dipartimento di Scienze economiche e metodi matematici

On Statistics

Corso di Laurea Triennale in Scienze statistiche

Lingua inglese

prof.ssa Paola Gaudio

A Selection from

Online Statistics Education: An Interactive Multimedia Course of Study

Developed by Rice University (Lead Developer), University of Houston Clear Lake, and Tufts University

Table of contents

http://onlinestatbook.com/2/index.html

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 11, 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 are 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.) 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.

Measures of Central Tendency

by David M. Lane

Prerequisites
• 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:

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

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:

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

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.

$$\mu = \frac{\sum X}{N} = \frac{634}{31} = 20.4516$$

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:

$$\frac{(4+7)}{2} = 5.5$$

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

Measures of Variability

by David M. Lane

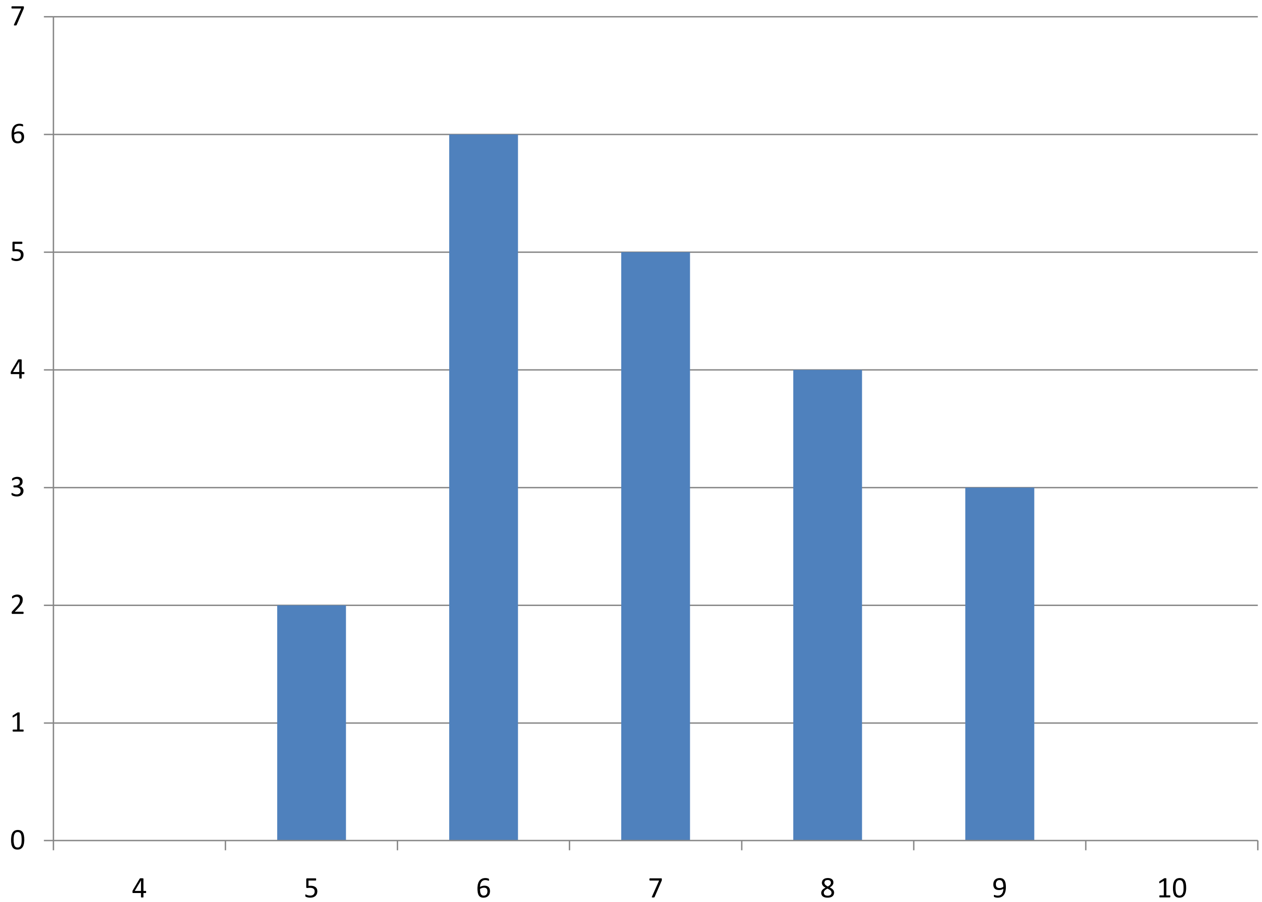
Prerequisites
- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: Measures of Central Tendency

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

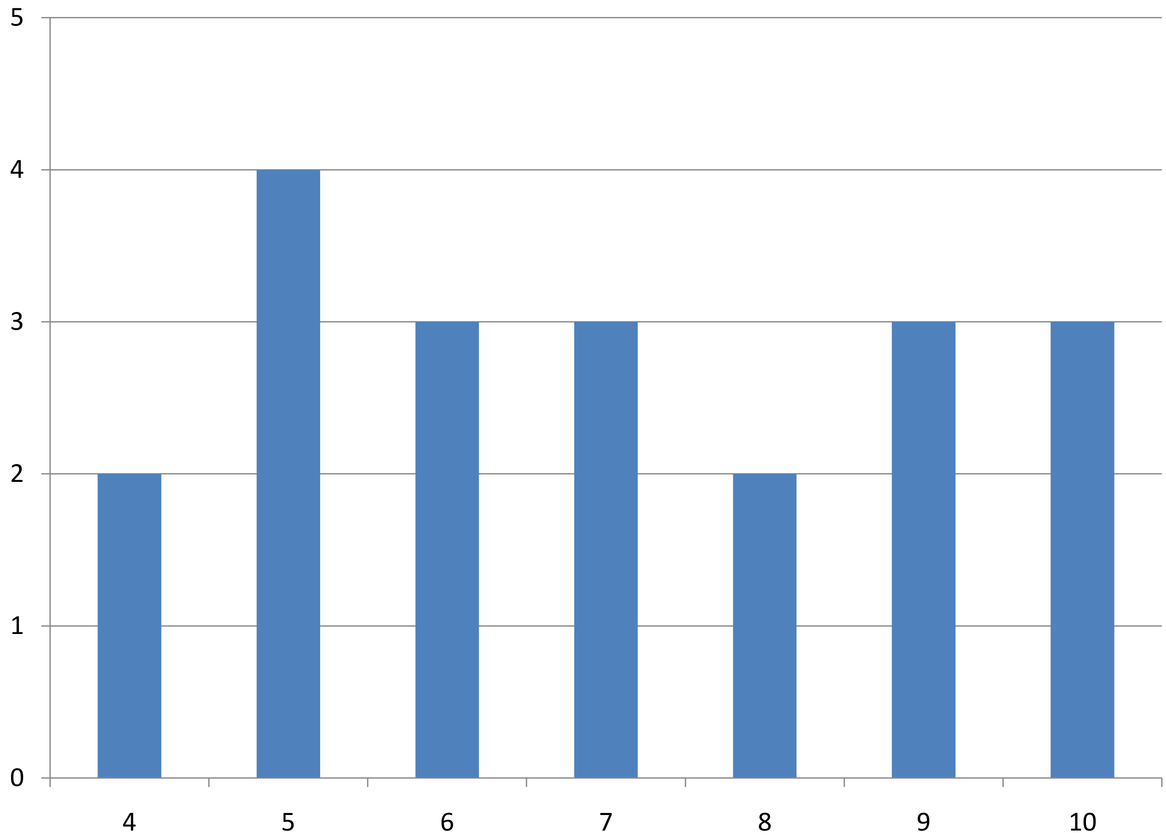


Figure 1. Bar charts of two quizzes.

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 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, μ = 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.

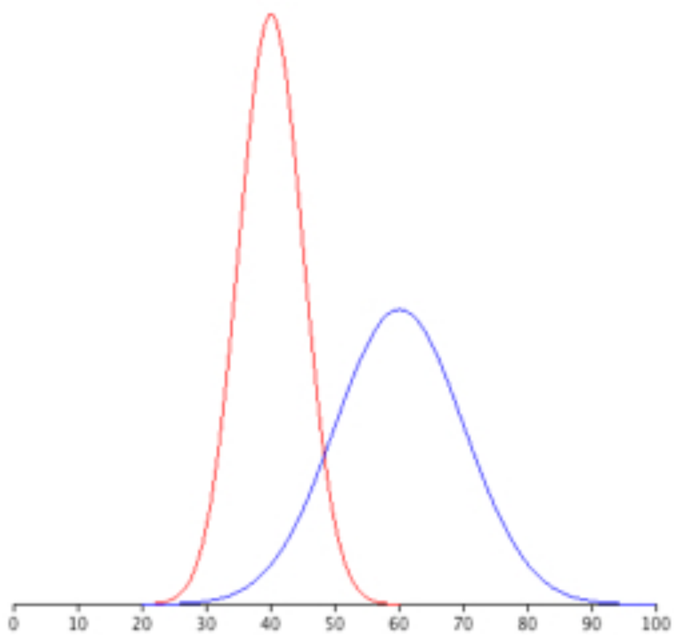


Figure 2. Normal distributions with standard deviations of 5 and 10.

Introduction to Normal Distributions

by David M. Lane

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

Learning Objectives
1. Describe the shape of normal distributions
2. State 7 features of normal distributions

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

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

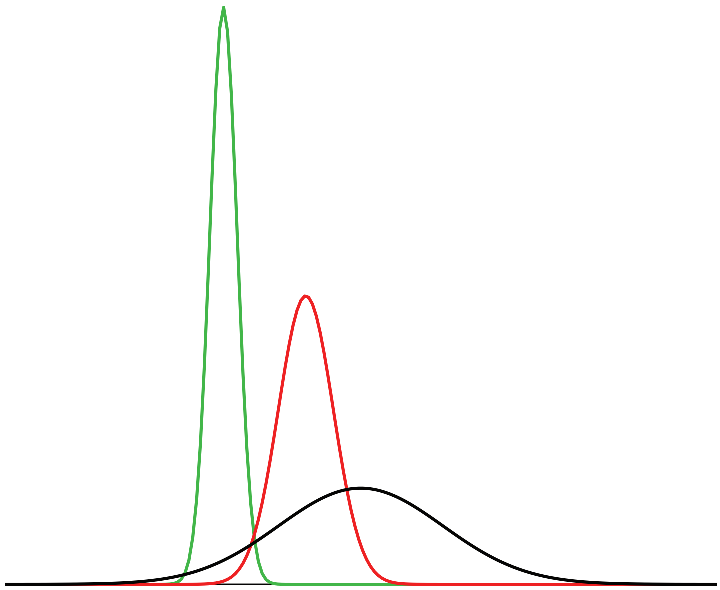


Figure 1. Normal distributions differing in mean and standard deviation.

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

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

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

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

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

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