PDHonline Course G392 (3 PDH)

Lies, Damned Lies, and Statistics

Instructor: Frederic G. Snider, RPG and Michelle B. Snider, PhD

2020

PDH Online | PDH Center

5272 Meadow Estates Drive
Fairfax, VA 22030-6658
Phone: 703-988-0088
www.PDHonline.com

An Approved Continuing Education Provider

Lies, Damned Lies, and Statistics

Introduction

The title of this course “Lies, Damned Lies, and Statistics’ is part of a famous quote by Mark Twain, who reportedly stated: "There are three kinds of lies: lies, damned lies and statistics." This quote refers to how statistics are commonly misapplied, often on purpose, to support a position or push an agenda. Statistics at its mathematical roots, however, has no such nefarious underpinnings, but is rather a way for us to grasp and communicate patterns and relationships within large sets of data without having to struggle with the data sets themselves.

Statistics is the study of data. A large collection of information by itself is difficult to for our brains to process, often leading to conclusions that can be at best meaningless and at worst misleading. Statistics is a mode of reasoning, a way of “mathematizing” data into a concise picture. It allows us to put information into a context, and gives us a way to discern its global behavior. Statistics are used in almost all human fields of endeavor, including:

- Politics: polls on opinions, how people will vote
- Education: assessment of course work, teacher and student performance
- Sociology: how is happiness related to wealth? How much tv do people watch?
- Sports: records, streaks, betting
- Art: how much things sell for, how much actors get paid, how popular a film is
- Medicine: How to decide whether or not to take a drug
- Science: How to interpret results of experiments.

Let’s start by defining what we are actually studying:

Definition: DATA- a collection of information represented as numbers.

Definition: A STATISTIC is a number that is derived from a set of data.

During this course, we will examine a variety of statistical methods that summarize sets of data to convey meaning. Each method gives some information about our data, but we will see that one statistic or even several statistics may not show us the whole picture.

In Chapter 1 we will look at statistical analyses when we have all the available information. In Chapter 2, we will look at the case where we only have some of the data. Finally, in Chapter 3, we will discuss some very interesting but little known aspects of statistics.

Chapter 1 – When We Have All the Information

In this chapter, we consider the situation in which we have access to all (or at least most of) the data for a given situation. First, we will look at a class of statistics called Single-Value Statistics, stats that use one number to represent a key fact about the entire data set.

Single-Value Statistics

The most commonly used single value statistic is the mean or average, defined as follows:

Definition: MEAN or AVERAGE: sum of all the values divided by the number of values.

The following list shows the final grades of 20 students in a college algebra class:

{71, 35, 67, 91, 85, 70, 75, 76, 90, 77, 78, 79, 99, 80, 81, 82, 86, 87, 70, 64}

To get the mean, add the numbers together and divide by 20. The sum is 1543, so the average is 1543/20 or about 77.

In statistic-speak, we use the variable $\bar{x}$, pronounced "x bar" to represent the mean. Restating the definition mathematically:

For any set of numbers, S={x_1, … , x_n}, the mean is given as

$$\bar{x} = \frac{\sum x_i}{n}$$

Where the ∑ symbol means "the sum of", and "n" is the number of items in the set.

Alternatively, we can order our list from lowest to highest, and find the value that occurs at the half-way point, defined as the median.

Definition: The MEDIAN of a set of data is the middle number, when data are listed in increasing order.

For an odd number of values, the median is just the middle number once the data is put in order. For example, the median of the data set {1,4,7,8,31} is 7.

If there are an even number of numbers, average the middle two. For example, for the data set {1,4,7,8,31,65}, the median is (7+8)/2 = 7.5.

Some more definitions:

Definition: The MINIMUM of a data set is the smallest value
Definition: The MAXIMUM of a data set is the largest value.

Definition: A QUARTILE is a quarter of the data points when the data set is listed in increasing order.

The FIRST QUARTILE represents the first 25% of the data, the SECOND QUARTILE gives 26% to 50% of the data, the THIRD QUARTILE gives 51% to 75%, and the LAST QUARTILE gives 76% to 100% of the data.

So for the grades listed above for the 20 students:

- The first quartile is {35, 64, 67, 70, 70}; for a range of 35 to 70
- The second quartile is {71, 75, 76, 77, 78}; for range of 71 to 78
- The third quartile is {79, 80, 81, 82, 85}; for a range of 79 to 85
- The fourth quartile is {86, 87, 90, 91, 99}; for a range of 86 to 99

Why would I want to do this, with the exception of giving the top 5 students in the class bragging rights? (“Na Na - I’m in the fourth quartile!”). Read on…

The Box Plot

A commonly used graphical way of displaying quartile data is called a **Box Plot.** The box plot uses the ranges of the first three quartiles plus the minimum and maximum giving us a 5-number summary of the data. First we draw a vertical axis that ranges from 0 to 100 (percent).

For our data set of class grades, the first quartile is the range 35-70. For this we draw a vertical line with a “T-bar” at the bottom.

The second quartile is 71-78, and the third quartile is 79-85. Draw boxes for each of these.

For the fourth quartile, draw a line for this range and put a “T-bar” at the top. The box on the left below shows the resulting box plot of this data set. By definition, half of the grades fall within the two boxes. The T-bars are at the minimum and maximum values.

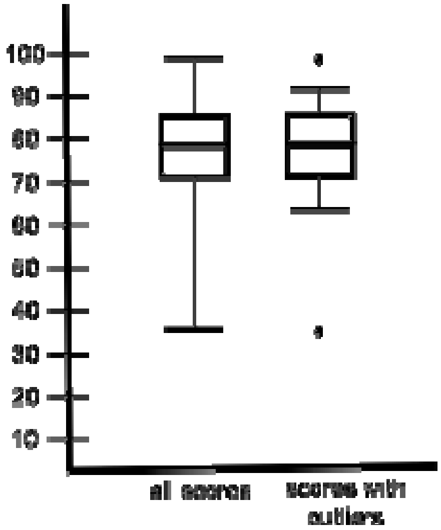


Figure 1 - The Box Plot

Looking back at the test scores, I see that the minimum score of 35 lies far below the second lowest score of 64. There is also a pretty big jump from the second highest score of 91 to the maximum score of 99. These two points lie outside the range of most of my data, so they get a special name - outlier.

Definition : An OUTLIER is a data point far from the rest.

To keep these two outlier students from affecting the box plot of the majority, I can separate the outliers, plot them as dots, and use the remaining data to make my box plot.

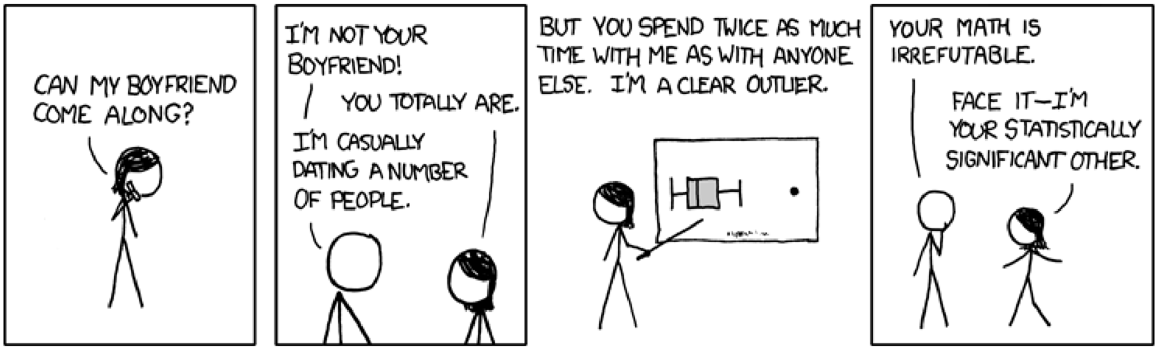
The plot on the right above is a box plot with the two outliers extracted from the data set, and the two outliers shown as dots at their scores. Note that the T-bars are much closer in than before, since we extracted the outliers and therefore changed the maximum and minimum.

The second box plot gives us a better view of the class results. A quick view of the box plot says: " Most of the students scored between about 60 and 90, and half the class scored between 71 and 88. One student failed miserably, and one totally aced it."

The definition of an outlier is purposely vague, as you have some discretion as to which points you consider outliers. Statistically, a good rule of thumb is to identify outliers as those values in the 2nd and 98th percentiles, meaning the lowest 2% of the values and highest 2% of the values. But this is always a judgment call.

But sometimes the truth is *in* the outliers, in this example:

In the 1970's, scientists conducted measurements of the thickness of the ozone layer in the upper stratosphere. It was hypothesized that the layer should be fairly uniform. Most of the data points were very close together, which seemed to support the hypothesis, but there were a few points near the South Pole which had very small measurements, close to 0. These were identified as outliers, perhaps equipment malfunction errors, and thrown out of the model. It was then concluded that the ozone layer was uniform. However, subsequent studies found that there was, in fact, a hole in the ozone layer above the South Pole. So the experimenters' original hypothesis and bias caused them to draw an incorrect conclusion from the data by labeling the unanticipated results as outliers.


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

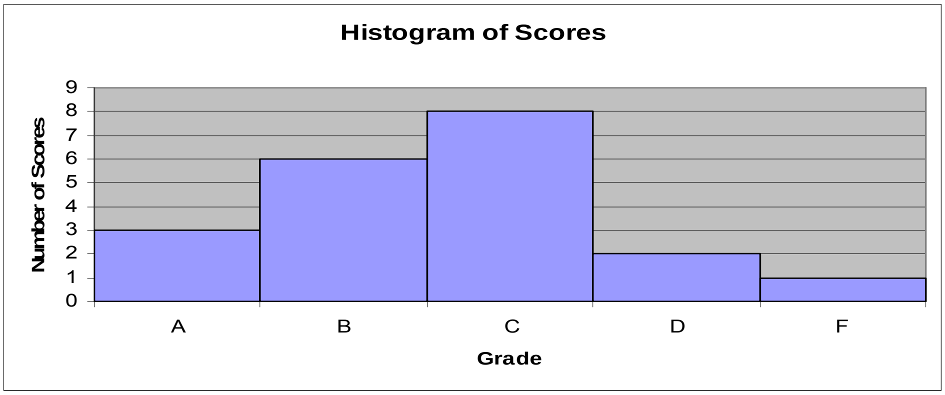
(http://xkcd.com/539/)

Histograms

Let's look at another way to present our data graphically.

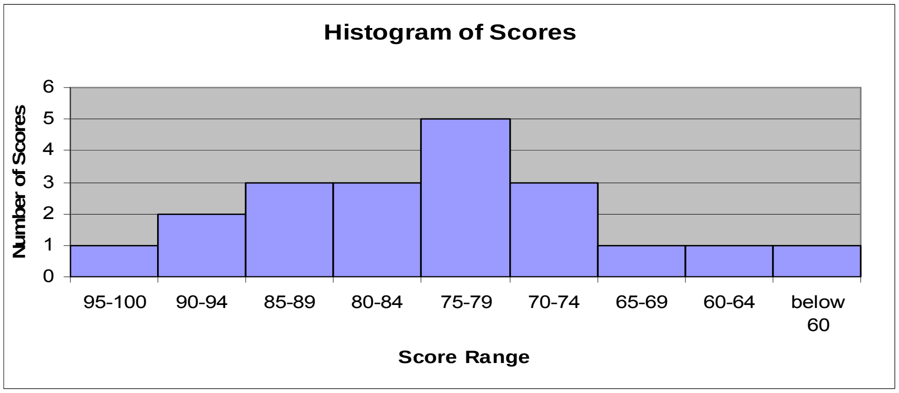
> **Definition: A HISTOGRAM is a bar graph, where the height of each bar is the frequency of occurrence of a data point or data range.**

The shape of a histogram will give more of a sense of the distribution of the data than we can get from a single-value summary or even a box chart. First, let's look at a histogram that shows the scores of the sample algebra students grouped by letter grade:



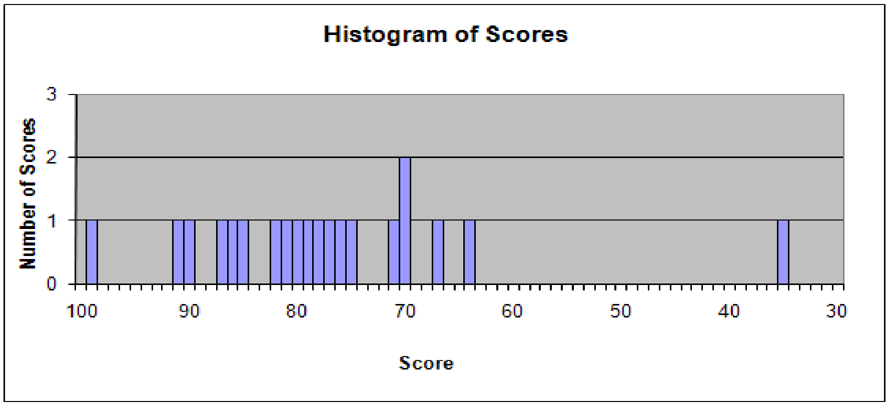
The histogram illustrates that 3 people received an A, 6 received a B, 8 a C, 2 a D and one an F.

The following histogram shows how the chart changes for the same grade set using 5-point score ranges:



This histogram indicates that one person received a grade between 100 and 95, two received grades between 94 and 90, etc. So this histogram is more detailed than the previous one, as we used a smaller score "window" for each bar. As we narrow the window, we get more bars and more detail.

In the extreme case, the next histogram plots the student's grade by each actual score. It shows us exactly how everyone scored, but no summary type information.

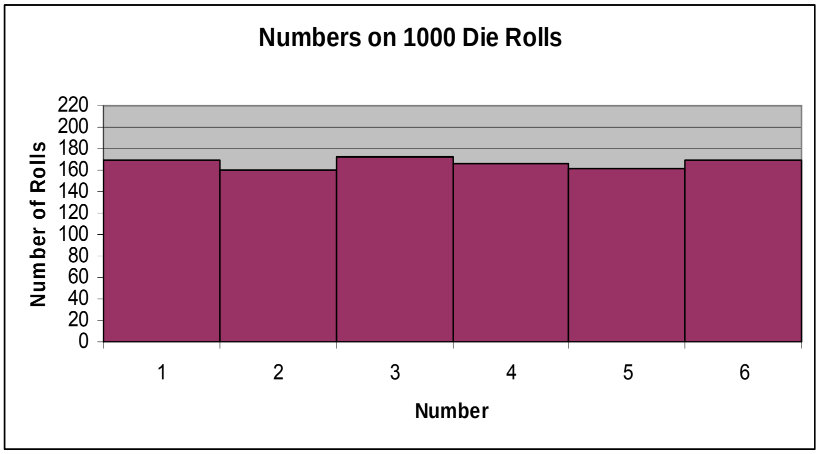


All the above histograms represent the same data set. From top to bottom, each histogram shows a little more detail at the expense of showing trends. For any data set, the level of detail which is the most useful depends on what you are trying to show. The lesson here is that the author of the plot greatly influences the reader's interpretation by selection of the window size.

Common Shapes of Histograms - Distributions

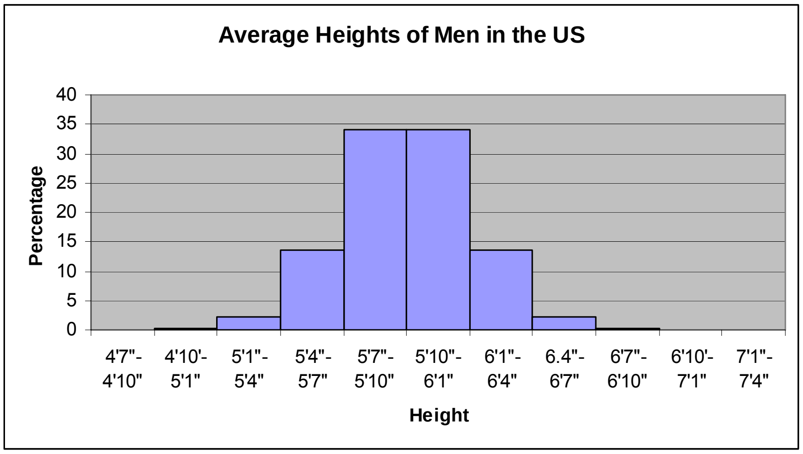
The shape of the histogram for a data set can in theory be anything. However, in practice, data sets generally fall into a limited number of categories, called DISTRIBUTIONS.

The simplest distribution is just a flat line, called a **UNIFORM DISTRIBUTION**. For example, if I roll a die 1,000 times, I would expect to roll each number about 1/6 of the time. Then my histogram bars have the same height for each value. In practice, 1/6 of the time corresponds to 166 2/3. Of course I can't roll a number a fractional number of times, so the histogram won't be a perfect straight line, but it might well look something like this:



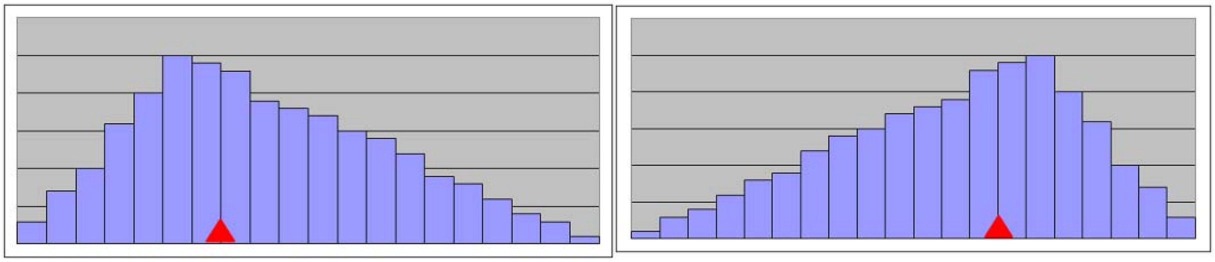
Example of a Uniform Distribution

A **SYMMETRIC DISTRIBUTION** has a roughly symmetric shape. For example, a histogram of the average heights of American males is shown below. The plot is the same shape on either side of the mean of 5’10” and so is a symmetrical distribution.



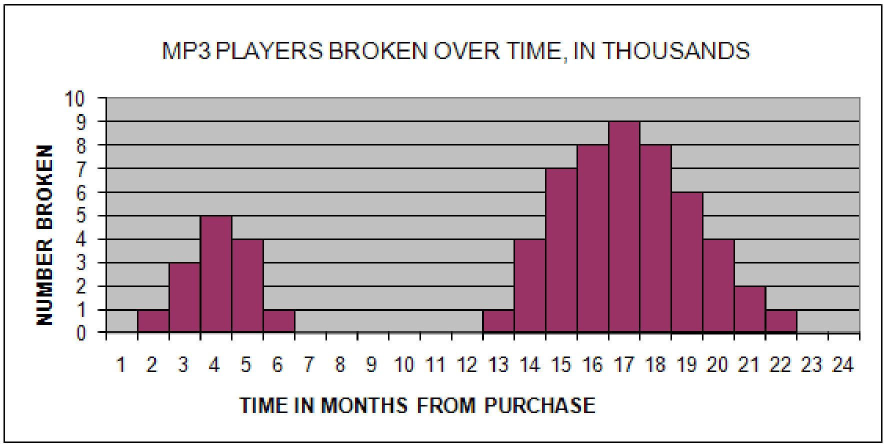
Example of a Symmetric Distribution

A **SKEWED DISTRIBUTION** has more data on one side of the mean than the other. We say a data set is “skewed to the right” if there is more data on the right side (tail is on the right), and “skewed to the left” if there is more data on the left (tail is on the left). Examples could include family income (skewed by a few very wealthy families), or per capita mortality versus age (more people die old than die young). On the follow two examples, the mean is shown by a triangle.



Example of a Right-Skewed Distribution and a Left-Skewed Distribution

A **BIMODAL DISTRIBUTION** has two peaks. Say I study 4,000 MP3 players of a particular brand, to see how long they will last. Manufacturing defects show up within a short period of time, and wear and tear issues usually show up later. It is in the company’s interest to know what this distribution looks like, so that they can sell you an extended warranty that will expire just at the “right time”. Here is a hypothetical distribution of breakage with time:

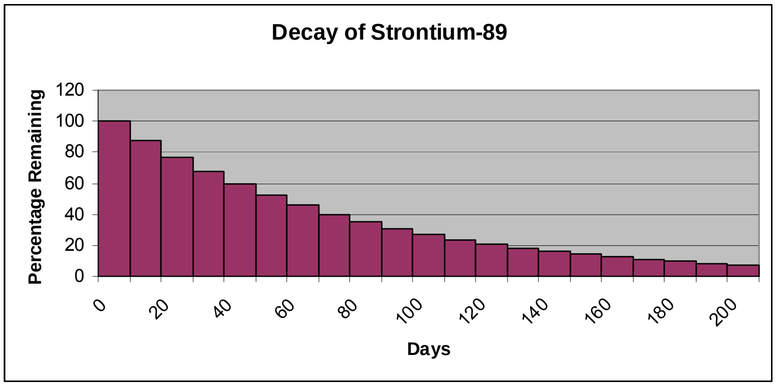


Example of a Bimodal Distribution

Usually the standard warranty will cover you for a few months, to cover the manufacturing defects. Based on the distribution shown on the histogram, it is in the manufacturer's interest to have the extended warranty expire at 12 months for this particular MP3 player.

An **EXPONENTIAL DISTRIBUTION** drops rapidly at the beginning, then slowly approaches (but never reaches) zero. The classic example is radioactive decay, where in a set amount of time called the HALF-LIFE, the quantity of material remaining drops by half. For example, Strontium-89 has a half-life of 53 days. Then, the percentage of the original quantity left after t days is given by,

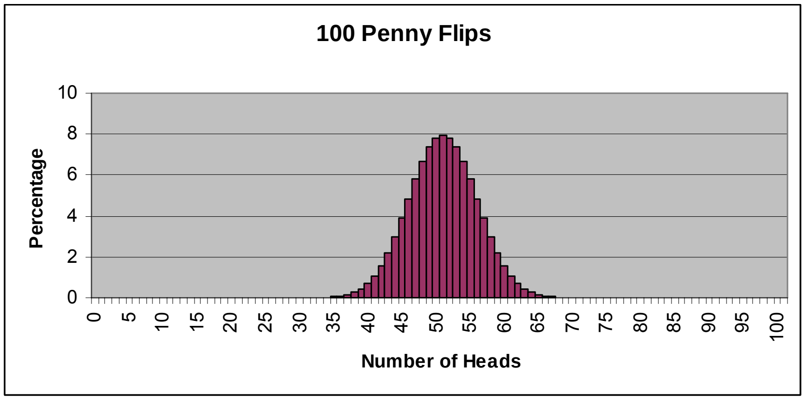
$$Q(t) = e^{-t\frac{\ln 2}{53}}$$



Example of an Exponential Distribution

A **NORMAL or GAUSSIAN DISTRIBUTION** describes a probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p.

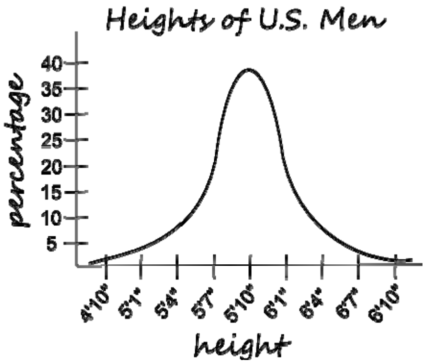
For example, flip a penny 100 times in a row and record the number of heads. Repeat this experiment many times. I can expect to get the following distribution of outcomes:



Example of a Normal or Gaussian Distribution

The histogram tells us that most of the time we will get between 40 and 60 heads, and only rarely more or fewer.

Often, instead of plotting the histogram, we plot the points at the top of each bar and fit a smooth curve to them. For the average heights of men in the U.S., we get the following plot.



A Normal or Gaussian Distribution drawn as a Bell Curve

You might recognize this distribution as a “bell curve”. We will discuss the bell curve in a minute, but first we have to digress to discuss the concept of the standard deviation.

Standard Deviation

We have talked about the mean of a data set. Now let’s look at a way to quantify how well the entire data set is clustered around the mean. Mathematically, we can calculate how far each data point is from the mean, then we can average these distances to get a sense of how spread out the data is overall. If we do that, the resulting number is called the Standard Deviation.

Definition: The STANDARD DEVIATION is the square root of the sum of the squared distances from the mean divided by the number of data points. It is represented by the Greek symbol sigma, and is written like this:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Where *x-bar* is the mean, *Xi* is each data point, and *n* is the number of data points.

For example, take a data set of four numbers: {-6, -4, 4, 6}.

Calculate the mean:

$$\bar{x} = \frac{(-6) + (-4) + 4 + 6}{4} = 0$$

Then calculate the standard deviation:

$$\sigma = \sqrt{\frac{(6 - 0)^2 + (4 - 0)^2 + (4 - 0)^2 + (6 - 0)^2}{4}} = \sqrt{26} = 5.1$$

(Sidebar: Often, the definition of the standard deviation has division by n-1 instead of n. That is the correct definition in the situation where we just have data about a sample of the population, not the whole population. In our example, we have all the data so we use n.)

You may wonder why there are squares and a square root in this equation. Why not just take the distances to the mean and average them? Let’s go back to our example. If we take the distances to the mean (0) and average them, we get (6+4+4+6)/4=5. What if instead we took the distance to a different value, like 1? Then we would get (7+5+3+5)/4 = 20/4=5. That is, we get the same “average distance” even if we aren’t measuring from the mean! This is true because if we are truly near the mean, some of the distances will be negative and some positive. By using squares and square roots, the signs of the distances don’t matter and we get a proper result.

Gaussian Distributions (The Bell Curve)

In the early 1800’s, the mathematician Carl Freidrich Gauss (1777-1855) was taking observations of the asteroid Ceres before it went behind the sun. He predicted where to look for it when it came around the other side by studying past observations and fitting a curve to the data in such a way that it minimized errors. The shape of his results turns out to be quite common and easily described mathematically, and now bears his name.

Definition: The Bell Curve, also called a NORMAL or GAUSSIAN distribution, is a curve that is completely determined by two pieces of information: the mean and the standard deviation.

The formula for the Gaussian distribution is given as:

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

where μ is the mean, σ is the standard deviation, and π=3.14159… and e=2.71828, which are both physical constants.

The notoriety of the Bell Curve comes perhaps from its ubiquity. It shows up in the average heights of American males and females, batting averages, and many other places. When college professors are assigning grades to their students, they like to do it in such a way that the scores form a bell curve, centered at their desired mean grade (like C+).

Let's see why this shape is so common. Adolphe Jacques Quetelet (1796-1874) was the first to apply Gauss' results to other situations. He studied what he called "l'homme moyen," the average man. He saw that men are distributed normally in attributes such as height, chest size, body mass index, intelligence, etc. Each of these descriptors is a result of many influences, genetic and environmental, that over a large group of people will tend to average out to a central value. We expect the bell shape, since most men are in a fairly small height range, with some a little taller or shorter, and fewer much taller or shorter.

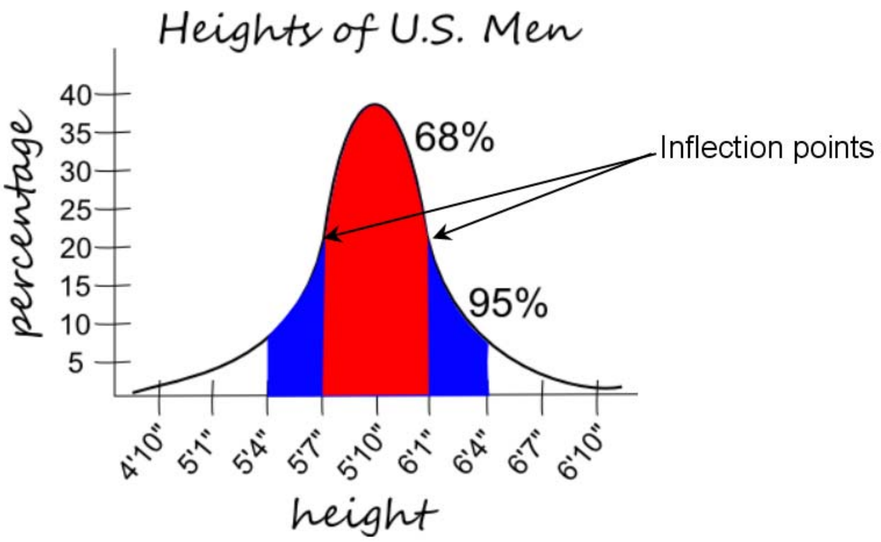
Think of it this way. Gaussian distributions arise when external influences are equally likely to contribute to the plus or minus side of the mean, so that the resulting graph is symmetric. This gives us our explanation of why the Gaussian distribution is so common: real-world quantities are often the balanced sum of many unobserved random events. For example, the binomial distribution (in cases such as flipping a coin a large number of times) is very close to a normal distribution.

The Gaussian distribution is so common across so many types of data, that this observation has been formalized into the following theorem:

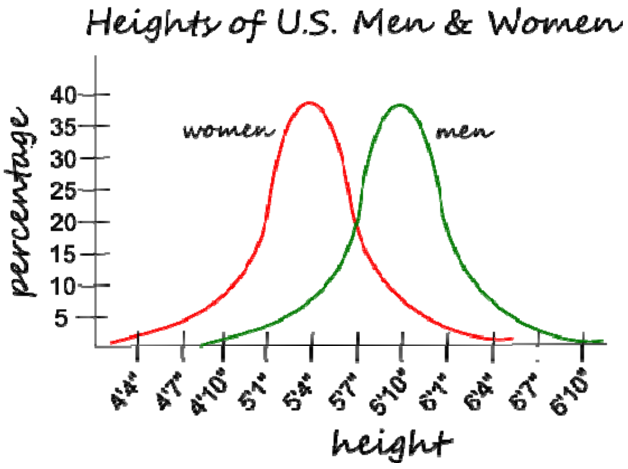
> **The CENTRAL LIMIT THEOREM: if we take a sufficiently large number of independent random variables, each with finite mean and variance, then we will get a normal distribution of outcomes.**

Notice that for small and for large values, the shape of a bell curve is concave up, and for values near the mean, the curve is concave down. The place where the graph changes from concave up to concave down is called the inflection point. There are two on every bell curve, one above the mean and one below.

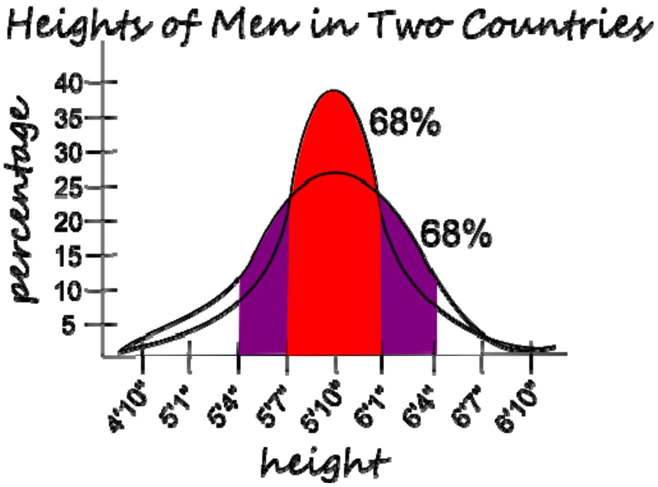
The inflection point corresponds exactly to the point one standard deviation below and above the mean. We can show mathematically that 68% of the area under the curve lies between these two points. The area determined by two standard deviations above and below the mean covers 95% of the area, and three gets 99.7%. In our example of the heights of U.S. men, the mean is 5'9" with a standard deviation of 3", so 68% of men are between 5'6" and 6', and 95% within 5'3" and 6'3".



If we compare the graph of average male heights versus average female heights, the shapes would be the same because they have the same standard deviation (3"), but the center would be shifted since on average women are shorter than men (5'4" versus 5'10").



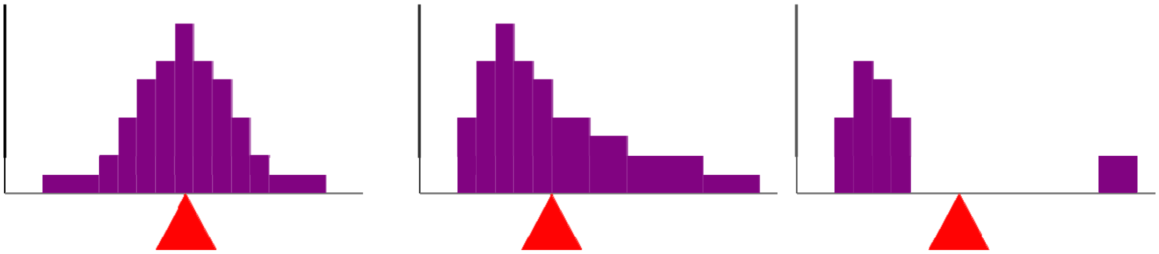
Now, say in another country, the average height of men is also 5'10", but the standard deviation is 6" instead of the 3" in the U.S. Then the graphs will look like this:



Notice that the area under the curve from one sigma above and below is still 68%. The standard deviation gives us a way to compare samples from different populations. For example, a man of height 6’4” is more rare in the U.S. then in our new country since he is 2 sigma from the mean in the U.S., but only 1 sigma from the mean in the other.

Problems with Single-Value Statistics

Although single-value statistics are useful, they do have some limitations. For example, let’s consider a physical interpretation of the mean. The three histograms below represent three data sets. Think of each one as equal-weight blocks on a see-saw. The mean is the place where you would put the fulcrum if you wanted the see-saw to exactly balance. Just as a small weight farther out on the see-saw would balance with a large weight closer in on the other side, a small number far from the mean is weighted the same as a large number close to the mean. In the following figures, the red triangle indicates both the balance point and the mean.



Histograms as blocks on a see-saw, with the mean as the pivot point

The upshot of this is that the mean is very sensitive to the influence of outliers. For example, if in addition to the test scores I had before, I had one student get a zero, the mean would now be given by

$$\overline{x} = \frac{1543}{21} = 73.5$$

This is 4.5 points lower than the previous mean of 77, which alters our conclusion about the class, with the addition of only one outlier. Knowing the mean salary for workers at Microsoft might not give you a good sense of how much the actual workers make: Bill Gates' salary is (probably) significantly higher and will skew the mean.

The mean can also be misleading in other situations. For example, according to the US Census Bureau, the average American has 1.83 children.

First of all, there is not a single person who has exactly 1.83 children. Secondly, there are multiple ways I can interpret this statistic:

- Scenario 1) Most people have either 1 or 2 children, with 2 being more common than 1.
- Scenario 2) About half the people have no children and about half have 4 children
- Scenario 3) About three-quarters of the people have no children and the rest have 8.
- Scenario 4) About seven-eights of the people have no children and the rest have 16.

All these scenarios give a mean of about 1.83 children. You would say, based on your own experience and personal observation, that scenario 1 is the correct one. But you can't tell that from the mean. What if I was describing a data set you knew nothing about? Then any of those scenarios is equally plausible.

To overcome the limitations of single-value statistics, we jump to multi-valued statistics.

Multi-Valued Statistics

Multi-Valued statistics are used when we have two or more data sets for a given population.

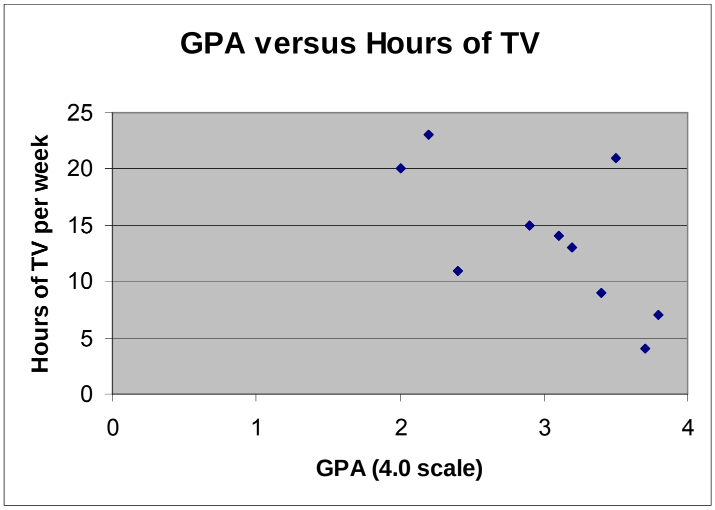
Two Data Sets: Correlation

Often, we will be presented with two sets of data for a given population, and asked to find if there is a relationship between them. Statistics gives us a way to analyze the data, to see if there is a pattern that lets us predict one from the other.

> **Definition: A CORRELATION between two data sets is a relationship between the two pieces of information.**

If we have two pieces of data for each member of a population, we can graph both data sets on the same plot. This can show us if the two pieces of information are correlated.

For example, we could look at the relationship between ten high school students' GPA and how many hours of TV they watch each week. On the following plot, each student is represented by one point on the chart, placed at an x-value of the GPA and a y-value of the number of hours of TV watched. Based on the plot, it looks like, in general, the higher GPAs are associated with lower numbers of hours of TV, but the points do not form a perfect line.



We can describe to what extent the quantities are moving together.

Definition: Two sets of data have a POSITIVE CORRELATION if one quantity increases as the other increases, and a NEGATIVE CORRELATION if one quantity decreases as the other increases.

For example, there is a positive correlation between high school SAT scores and college GPAs, but a negative correlation between the life expectancy of a person and the number of cigarettes he/she smokes a day.

To determine how strong the correlation is, we first compute the mean and standard deviation of the two data sets individually, as we have done before. Then, we compare data points with the same standard deviations.

In our example, the mean GPA is 3.02, with standard deviation of 0.63. The average number of hours of TV watched is 13.7, with a standard deviation of 6.23. Now we can compare: if I pick a person whose GPA is 1 standard deviation above the mean, is his number of hours of TV also one standard deviation above the mean as well? If so, then we say the data is correlated. If these calculations correspond exactly for all of data points, we get the following special case:

Definition: A data set is PERFECTLY CORRELATED if we can draw a straight line on the scatter plot that goes through all of the points.

Let's be more precise. We mathematically define the correlation as the following:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad \text{and} \quad s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

The expression $\left(\frac{x_i - \bar{x}}{s_x}\right)$ is how many standard deviations away you are from the mean, which is also called the *z*-score.

In general, using these formulas, the correlation will be between -1 and +1. Exact positive correlation gives +1, and exact negative correlation a -1. A correlation of 0 means the two pieces of information are apparently not related at all.

Given our scatter plot, we can draw a line that approximates the direction of the data.

> **Definition: the REGRESSION LINE is the line that most closely approximates the direction of the data.**

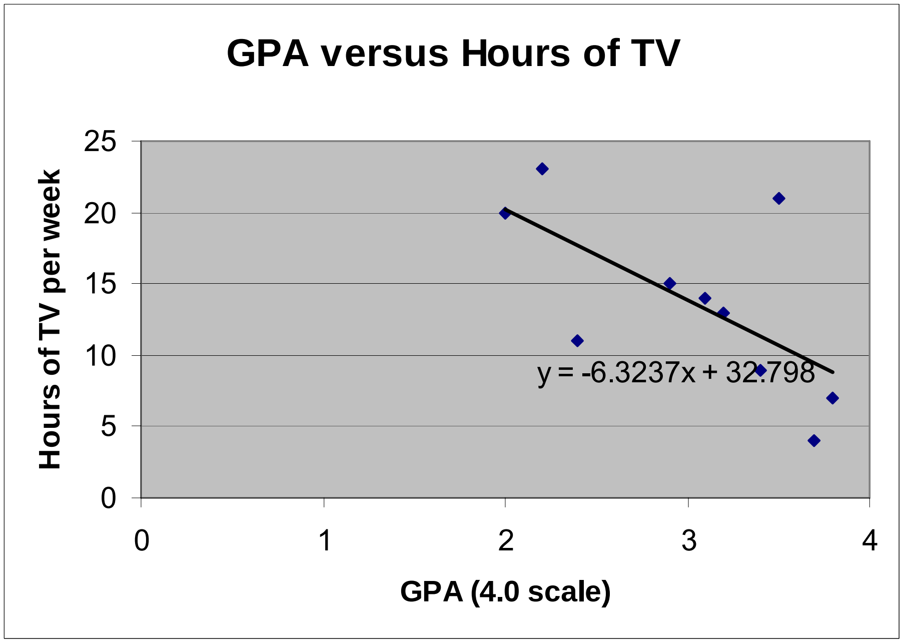
> **Definition: The RESIDUAL is defined as the vertical distance between a given point and the regression line.**

The best we can do is to try to get the line as close to as many data points as we can. Mathematically, this is given by the following line:

> **Definition: the LEAST SQUARES REGRESSION LINE is the line that minimizes the sums of the squares of the residuals.**

Let's find the regression line for our graph (you can do this in most spreadsheet programs like Microsoft Excel ©.)

$$y = -6.32x + 32.8$$



Now that I have the regression formula Y=-6.3237x + 32.798, I can tell how many hours of TV you watch if you'll tell me your GPA. Say your GPA is 2.5. Plugging that in for x gives:

Y= - 6.3237 (2.5) + 32.798, which is16.9886 hours of TV per week.

Even though my answer comes out to the forth decimal place, note that the scatter in the original data is large. So I will need to caveat my answer with a pretty wide error range. Just by eyeballing the graph, I would probably guess that if your GPA is 2.5 you probably watch between about 12 and 22 hours of TV per week.

The lesson here? Beware of coefficients that imply more accuracy than the basic data can support.

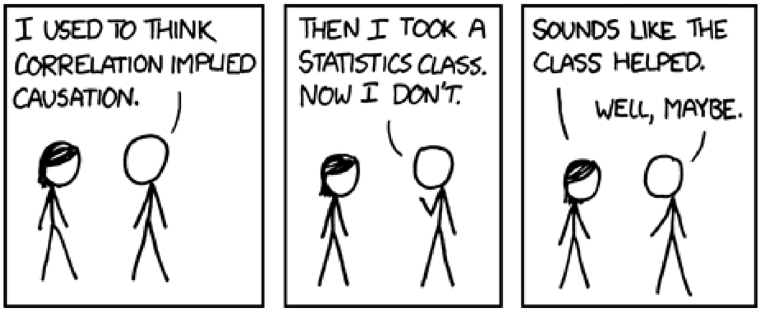
Correlation Versus Causation

News Flash: Local Woman Controls the Weather!

Orlando. This week, there was a 50% chance of rain each day. Monday I forgot my umbrella, and it rained. Then Tuesday through Friday, I remembered my umbrella and it did not rain. Could I therefore conclude that the act of bringing my umbrella actually prevented the rain from happening? There is definite correlation. Common sense, however, says that one is NOT the result of the other.

Definition: CAUSATION is when the change in one variable is the direct result of the change in another.

In his manifesto, "In Defense of Food," Michael Pollan points out that there is a correlation between taking vitamins and overall health, but that there is not necessarily a causation, as people who take vitamins tend to be more health-conscious than those who don't, and thus maintain a healthier lifestyle. There is a correlation between the two, but that does not mean that taking vitamins causes better health. Mathematicians can address correlation. Causation is outside the realm of statistics.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

(http://xkcd.com/552/)

Blurring the distinction between correlation and causation is one of the most common ways to manipulate data to support a position or further an agenda. It is important that we all understand the difference.

More Than Two Data Sets: Multiple Regression

The concepts above can be extended to cases where we have a collection of variables that may or may not correlate to the variable of interest. For example, say we are interested in analyzing home prices in a particular city. It is likely that we may need to look at variables such as number of bedrooms, number of bathrooms, square footage, distance from city center, lot size, age of home, crime rate, and so on, as all of these factors will affect the price of a house.

Programs like Excel© or specialized plotting programs can easily create the two dimensional graphs like our GPA example above, and even make 3-D graphs which would allow visualizing the relationship of three variables instead of two. Using various colors on a 3-D graph can potentially help us visualize four variables. More than four, however, and visualization becomes very difficult.

With our home-price analysis example, we have at least 7 variables. We could treat them in pairs, like home price vs. number of bedrooms, home price vs. number of baths, home price vs. square footage, etc. and make a bunch of plots and see if that helps us understand the data. Sometimes that works, at least to determine positive and negative correlations. But that is not a holistic approach and tells us nothing about how the variables interact with each other.

As it turns out, the mathematics to determine the correlation between two variables discussed above scales up to as many variables as we need without a hitch. These days you don't need a PhD in math to do it - you just need the right software, and an understanding of how to input the data and interpret the results. The mathematics is called multiple regression since we are looking at how multiple variables relate to the variable of interest. The details are beyond the scope of this course – suffice it to say that multiple regression is a tried and true statistical tool.

Chapter 2: When We Don't Have All the Information

There are many cases where we want to an analysis but we don't have (or can't get) all the information. For example, if we wanted to know exactly how many people are watching a certain television show , we would need an army of people calling everyone in the country at the same time and asking them if they are watching our show. What is the median salary of office workers, or the number of people who ride bicycles to work? Collecting all the data to address these questions is certainly not feasible. So we have to take some sort of shortcut.

Statistical Inference

The main idea of statistical inference is to use information about some of the members of a group to draw conclusions about the group as a whole.

Definition: The POPULATION is the total collection considered.

For example, we could look at all eligible voters, all the students at a particular university, or all the auto parts produced in a year. Note that we use the word “population” for the data set, even if it does not consist of actual people. We will consider the case where we can’t get data for the whole population; just for some members.

Definition: A SAMPLE is a subset of the whole population.

A common example of this is political polling before an election, where we ask some people how they will vote, and try to predict the outcome for the country as a whole. This is also common in test marketing, where a company will try to gauge how well a product will do by extrapolating from the feedback of small group of people. It is important to choose a sample well. Depending on what information you are gathering, you might want to focus on a particular demographic group. If I’m doing a marketing survey, I would like to poll people who would be interested in my product.

The key to statistical inference is sample randomness. For example, if I poll only Democrats, I will have a very different election prediction than if I poll only Republicans. Ideally the characteristics of my sample will reflect the characteristics of the population as a whole. Of course, the tricky part is that I don’t know the characteristics of the population as a whole. If I am doing an (unbiased) political poll for an upcoming election, I want to be sure to sample a range of ages, income levels, ethnic groups, and political leanings.

If the experimenter is allowed pick people to poll in certain groups, he may bring his own biases to the selection process. The goal is to have the sample’s characteristics be representative of the entire population. So, what we want to know is, if I take a certain sample, how close are the shape, center, and spread of the data to what I would get if I had all the information, and how sure can I be about my conclusions?

The most common way to choose a sample is randomly:

Definition : A SIMPLE RANDOM SAMPLE (SRS) is a sample taken at random from the population.

For example, I want to find out about the heights of adult men in the U.S. I can get an idea of the heights of the whole population by choosing a few representative samples, whose heights are indicative of the rest. Of course, this would be easy if I already knew what the average height was. So the challenge is to find a sample that will give us good information. The best way to do this is to choose men completely at random.

We can also assume that the histogram we will eventually get should be normally-shaped, based on our general observations about heights. So I pick my first man. I know that he is more likely to be near the mean, because in a normal distribution, most people are. Then, I take a second man and add his data point. The more people I add, the closer I will get to the actual mean, since it will become increasingly unlikely that I get only values at the extremes, and outliers will tend to cancel each other out by pure randomness. (That is, I am unlikely to only pick only extremely short men and normal men, and not extremely tall men.)

So, we take a sample, and then we assess the probability that we would get such a sample if the characteristics of the population were significantly different than the sample. For example, to test if a medication works, we can take a group of people with the disease, give half the medication and half a placebo, and then see if there is a difference in the number of people cured in each case. Then we can predict that the same ratio of patients cured will occur in the larger population.

Probability can help us determine if a difference is significant or within a reasonable range of expected values. There are a number of ways to address the issue of how representative a sample is compared to the entire data set, each with its own complexities and pitfalls. Unfortunately, these topics are outside the scope of this course. We say unfortunately only because the authors of this course love this stuff and love to talk about it.

Sampling Biases

Let's look at a classic example of biases in polling. In 1936, the two presidential candidates were (incumbent) Franklin D. Roosevelt and Alfred M. Landon. At the time, there was a magazine called the Literary Digest, which ran polls before major elections. They had been successful at predicting the winner on a few of the previous elections, so they had a good track record. They sent out 10 million surveys, and got 2.4 million replies. With this data, they predicted that Landon would not only win, but that it would be a landslide of 370 to 161 electoral votes. When the actual election results came out, Landon won only 8 electoral votes to Roosevelt's 523.

An analysis of why they were so off revealed three major biases in the sample. First, the mailing list was cobbled together from their subscriber list, car registration records, and telephone records. Notice that this was in the middle of the Great Depression, when cars and telephones were luxury items. When money is tight, magazine subscriptions tend to be one of the first things to go. So the responses were primarily from wealthy people, whose priorities and thus political leanings were not representative of the population at large.

Second, the survey was voluntary. Notice that only 1 in 4 people returned the survey, so there is a fair amount of self-selection in the data. The third major issue with polling is that people may not tell the truth. Maybe they are embarrassed, or maybe they are just undecided, or just say what their friends are saying.

These factors influence the data set such that you think you know what is going on, but in fact the data completely misrepresent the population as a whole.

Addressing Liars in Surveys

If the problem with your particular survey is that people have a tendency to lie, there is a way to design an experiment that accounts for this. Suppose I want to know how many of my students cheated on the last test, but I know that no one will confess since there would be consequences.

I ask the whole class to do the following: each student flips a coin (and doesn't show it to anyone). Then, I say "Raise your hand if you cheated OR you flipped heads." For any individual student, she may have flipped heads, or cheated, or both, but has not incriminated herself by raising her hand. So what does this tell me? Say I have 1000 students, and 800 raise their hands. I would expect 500 of the students to flip heads and 500 to flip tails, by pure probability, so I can fill out the following chart:

	Raised Hand	Did Not Raise Hand	Total

Flipped heads:	500	0	500
Flipped tails:	300	200	500
Total:	800	200	1000

Then of the students who got tails, 300/500 cheated, or 60% (ouch). Of course, this is assuming exactly 500 students got tails. If we look at the normal distribution of flipping 1000 coins, we get that it is quite likely that we get in the range of 460-540 students getting heads, which would give a percentage range of 57-63% of cheaters.

Chapter 3 - Interesting and Amazing Aspects of Statistics

Although most people think of statistics as dry and boring, there are some interesting aspects.

Lurking Variables – Science Fair Nemesis

When we do an analysis of a data set or design and experiment, we hope that we have thought through all the possible variables that could affect our results. Sometimes this isn't true. Let's look at some conclusions of recent experiments revealed at a high-school science fair:

- Conclusion: The size of a person's vocabulary is directly related to their foot size.
- Conclusion: Students with tutors do more poorly on tests than those who don't.
- Conclusion: Church attendance increases drinking.

Definition: LURKING VARIABLES, or confounding variables, are variables that affect both the independent and dependent variables, but may not be recognized by the experimenter.

If one is lucky, once the lurking or confounding variable is identified, the data can be re-evaluated with this new information and it will make more sense. Here are key lurking variables missed by our high school students:

- Conclusion: The size of a person's vocabulary is directly related to their foot size.
  (The age of the subject affects both of these in most cases.)
- Conclusion: Students with tutors do more poorly on tests than those who don't.
  (Students who do well academically don't need tutors!)
- Conclusion: Church attendance increases drinking.
  (Larger cities tend to have both more churches and more liquor stores.)

Simpson's Paradox: Two Contradictory Conclusions

Ms. Washington has applied to a college and gotten rejected. A total of 2000 applicants applied for 1100 openings in the freshman class. She is suing, claiming that the admissions practice favors men over women. Her lawyer presents the following data for the schools' admissions that year:

	Accepted	Rejected	Total	Acceptance Rate
Men:	700	300	1000	70%
Women:	400	600	1000	40%
Total:	1100	900	2000	55%

Note that we have eliminated all the variables except gender. The question is, if there was no bias and admission was completely gender-independent, how often would such a distribution occur for 2000 applicants, of whom 1100 are accepted?

If there were no bias, and 55% of the total applicants are accepted, then 55% of the women applicants should be accepted and 55% of the men should be accepted. However, as shown on the table, only 40% of the women were accepted, while a whopping 70% of the men were accepted. The prosecution argues that certainly there is gender bias going on!

The defense (the college) comes back with the following information. There are in fact two separate programs at the school, the Academic Scholars Program and the Standard Curriculum. Each program has its own set of acceptance criteria. The scholars program has 240 places for next year, and the standard curriculum as 860, for a total of 1100 total places, as above.

Breaking down the total admissions information into the two programs looks like this:

Scholars	Accepted	Rejected	Total	Acceptance Rate
Men:	40	160	200	20%
Women:	200	600	800	25%
Total:	240	760	1000	24%

Standard	Accepted	Rejected	Total	Acceptance Rate
Men:	660	140	800	82.5%
Women:	200	0	200	100%
Total:	860	140	1000	86%

Notice that in <u>both</u> subprograms, the acceptance rate is higher for women. In the scholars program, five times more women were accepted than men (200 women versus 40 men). Furthermore, every single one of the women who applied for the standard curriculum were accepted, while only a little over 80% of the men were accepted. So the college argued that, In fact, it looks like men are being discriminated against, not women!

> **Definition: SIMPSON'S PARADOX describes the case where under two different (valid) analyses, the same data support two contradictory conclusions.**

Regression to the Mean (Or, why I always get better when I make a doctor's appointment)

Recall that the mean of a set of data is an average, so that some data points will be above and some will be below. In particular, if there is a streak of values that are above the mean, it is likely that the streak will be followed by a drop in the values.

> **Definition: REGRESSION TO THE MEAN is the phenomenon in which if one data point is far from the mean, the following point will be more likely close to the mean.**

Say there is an intersection near my house at which there is an average of one accident per month. This month, there were three accidents. Then, it is quite likely that there will be fewer accidents in the next several months. This can cause false conclusions in some situations: perhaps after the accidents, the police install a red-light camera. Then, they could claim that the reduction in accidents was due to the fact that they have been averted by increasing the public awareness of the danger by introducing the camera, when in fact the camera had nothing to do with it.

This phenomenon appears in many different places. For example, if a movie star overall does well in about half of her films, but then does several successful films in a row and makes it onto the cover of Cosmo or People magazines, it's quite likely that her next film will bomb. Perhaps I come down with a cold and stay home sick for a week before I go to see the doctor, and then I get better the next day even though I haven't picked up my prescription yet. Tall people may have short children. Things will always tend to return to the average.

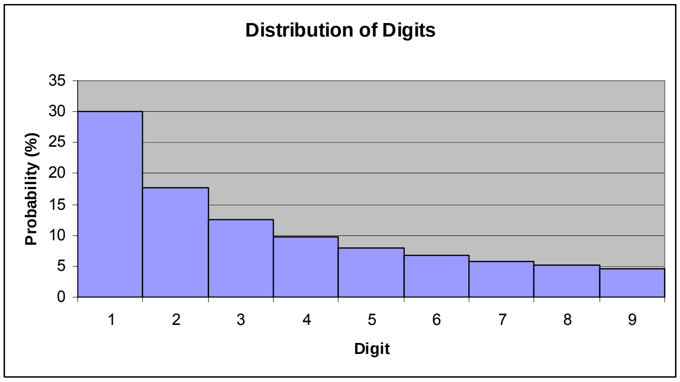
Benford's Law and Fraud Detection

Say I am doing an experiment, and I get a large set of data as a set of numbers. You might guess that if I look at the distribution of the first digits of that data, I would get a uniform distribution of leading digits. However, often, in practice, the distribution of digits is non-uniform in a very specific way. This is described by the following:

> **Definition: BENFORD'S LAW, also known as the First-Digit Law, states that the probability of the leading digit being the value d (for d= 1 to 9) is**

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(1+\frac{1}{d}\right)$$

Here is a histogram of the expected probability by first digit:



And here are the numerical probabilities:

Digit:	1	2	3	4	5	6	7	8	9
Probability:	30%	17.6%	12.5%	9.7%	7.6%	6.7%	5.8%	5.1%	4.6%

That is, the probability of getting a leading digit of 1 is 30%. This was first noticed by the American astronomer Simon Newcomb in 1881, who observed that the pages of the book of logarithms were much more worn towards the front of the text. He wrote a paper, which was ignored, and the phenomenon was then documented by Frank Albert Benford in 1938. It holds on data sets as diverse as street addresses, electrical bills, populations, death rates, and physical constants: Taking data from several disparate sources, the table below shows the distribution of first digits as compiled by Benford in his original paper.

Title: **Digit:**	**%1**	**%2**	**%3**	**%4**	**%5**	**%6**	**%7**	**%8**	**%9**	**samples**
Rivers, Area	31	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
Constants	41.3	14.4	4.8	8.6	10.6	5.8	1	2.9	10.6	104
Newspapers	30	18	12	10	8	6	6	5	5	100
Specific Heat	24	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
Drainage	27.1	23.9	13.8	12.6	8.2	5	5	2.5	1.9	159
Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
n^(-1), sqrt(n)	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8	8.9	5000
Design	26.8	14.8	14.3	7.5	8.3	8.4	7	7.3	5.6	560
Reader's Digest	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
X-Ray Volts	27.9	17.5	14.4	9	8.1	7.4	5.1	5.8	4.8	707
Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3	1458
Blackbody	31	17.3	14.1	8.7	6.6	7	5.2	4.7	5.4	1165

Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5	5	342
n^1, n^2...n!	25.3	16	12	10	8.5	8.8	6.8	7.1	5.5	900
Death Rate	27	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average	30.6	18.5	12.4	9.4	8	6.4	5.1	4.9	4.7	1011

(http://mathworld.wolfram.com/BenfordsLaw.html)

Let's look at an example where this makes intuitive sense. I have a bank account with 10% interest into which I invest $1. Then for the first couple of years, the interest will begin with a 1 and my nest egg will not grow very quickly. Then, once I hit higher principal, the interest will be higher, and the value of my investment will make larger jumps through the higher first digits, until I get to $10 and the process repeats. Here is how my investment grows at 10% interest:

$1.00, $1.10, $1.21, $1.33, $1.46, $1.61, $1.77, $1.95, $2.14, $2.36, $2.59, $2.85, $3.14, $3.45, $3.80, $4.18, $4.59, $5.05, $5.56, $6.12, $6.73, $7.40, $8.14, $8.95, $9.85……

There are 25 numbers here, 8 of which have a leading digit of 1, or about 30%, just as Benford said.

This phenomenon makes sense in the context of exponential models as above. However, it appears in many non-exponential contexts as well.

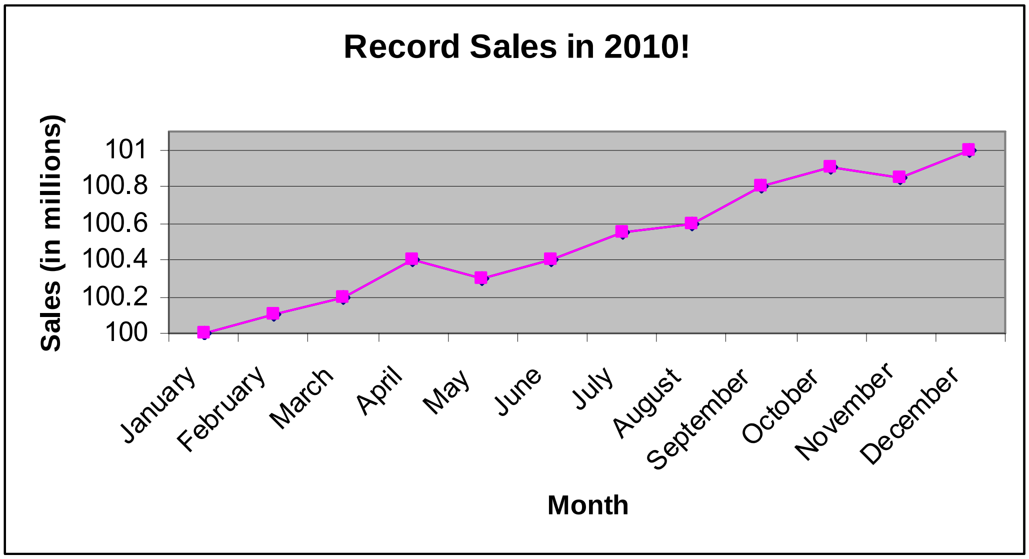
Benford's Law is useful in detecting fraud, since people who are fudging data that they want to look random will tend to make a much more uniform distribution. In particular, people tend to include disproportionately more 5's and 6's because they are "average" numbers.

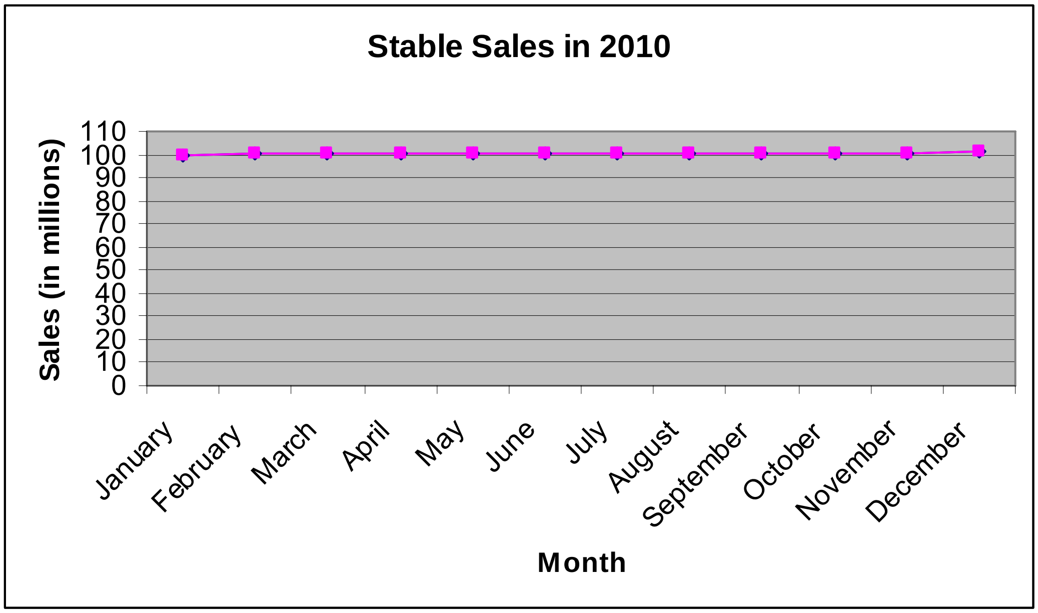
The Numbers Never Lie, but They Can Mislead!

We've seen that it is rare that a single value gives a good representation of a data set, so we need to keep this in mind when we are presented with statistical data. We have already discussed the perils of correlation versus causation, and the possible presence of lurking variables. Now we will look at a few other ways in which statistics can be misleading.

Reframing the Same Data

Graphs can be rescaled to make data give the impression of a desired conclusion. Consider the following graphs:





It looks like they are very different, but in fact they are the same data!

This often comes up in political races, where the same data is analyzed to support opposing conclusions. For example, say there is a proposed tax cut that will save every person about 3% of their taxes. The candidate who is trying to win votes from the middle classes will emphasize the equality of the cuts, saying that everyone is paying about the same. The opposing candidate perhaps can win votes from the higher income brackets by instead showing a straight up monetary cost of the tax cuts, showing that in terms of absolute money, they will save a lot more than people who make less. Both candidates are telling the truth, but with different spins.

The Perils of Percentages

When screening candidates for high security government positions, security officials often perform polygraph tests. Studies have shown that these tests are not 100% accurate: people who are lying can sometimes pass the exam. Perhaps a government official justifies their use by saying, “The exam is 80% accurate in detecting spies.” What does that actually mean? Let’s say the exam is performed on 10,000 potential employees, 10 of whom are actually spies. Then, with an 80% success rate, the test would correctly identify only 8 of the 10 spies. Furthermore, 20% of the non-spies would also fail the test, or 1,992 people. Of the 1992+8 = 2000 people failing the exam, 1992/2000 = 99.6% of them are not spies! Because of how often an innocent person can fail and a guilty person can pass, polygraph results are not admissible in courts, although they are still required for government clearances. (In practice, they serve more to deter espionage than to actually detect it.)

But what is your basis for comparison?

People are often misled because they don’t have a good sense of scale for large or small numbers, and numbers without comparison are hard to judge. Say I have planned a trip to Israel. I call up my mother to let her know, and she says, “That is unsafe! You have a good chance of dying in a terrorist attack! Last year 248 people were killed!” Upon further investigation, I determine that the population in Israel is 6.5 million people, so people killed by terrorist attacks constitute only 0.000038% of the population. For comparison, there are about 42,000 traffic accident deaths per year in the U.S., where the population is 290 million, which gives a rate that is almost 4 times higher! That doesn’t make me stop driving.

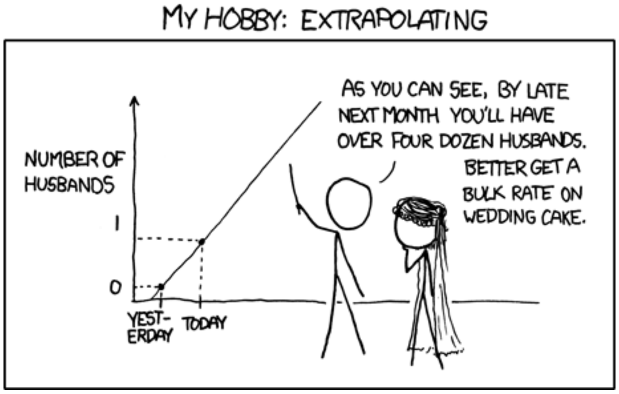
Say a drug company claims that taking their drug will reduce your risk of a heart attack by 30%. That sounds fantastic, right? But perhaps your original chance of dying of a heart attack was about 1.3 out of 10 million. Then, if their claim holds, you will now only have a 1 in 10 million chance. Is that worth the side effects? Maybe I run a large company, and I’m concerned about our stock prices. In our millions of dollars of business deals, we ended up last year with only a $1000 profit. This year, we made $1500. I can then issue a press release that says we did 50% better than last year! It’s not wrong, but it is a misrepresentation of the state of the company.

In 1995, O.J. Simpson was prosecuted for the murders of his wife Nicole Brown Simpson and Ronald Goldman. The prosecution presented evidence including proof that on previous occasions he had beaten his wife, and transcripts of 9-1-1 calls. The defense then cited the statistic that only 1/1,000 men who beat their wives go on to murder them, in order to lead the jury to think it was very unlikely that Mr. Simpson would have done it. What they did not say is how many men who do NOT beat their wives go on to murder them! Even if 1/1,000,000 men in the general population murder their wives, that means that a man who beats his wife is 1,000 time more likely to kill her. Strategies like this are often employed by lawyers to obfuscate the evidence and confuse jurors.

Another issue that came up in this trial had to do with DNA evidence. Say there was a particular characteristic in a DNA sample taken at the scene of the crime, and only 1 in 1 million people have that characteristic. Then the prosecution can cite this as evidence that the defendant is guilty. However, it could be the case that that evidence was used to find the defendant in the first place. For example, the police could have typed the DNA, then searched a database of the whole city for people who had that DNA type, found only 100 people, and chosen a suspect from that list. Then there is really only a 1/100 chance that that particular person is guilty, and citing the statistic that only 1 in a million people have this type of DNA is a flawed argument. There is bias in the sample.

Dangers of Extrapolation

If we are using data to predict future behavior, we have to take into account all the different factors, lest we draw inappropriate conclusions. Think about a turkey on a farm. Every day he is fed and taken care of, for 364 days straight. He may conclude that he will continue to be safe, as the data supports that prediction. However, the next day might be Thanksgiving.


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

http://xkcd.com/605/

The Future of Statistics

Statistics plays a key role in our understanding of the physical world. We collect data and then analyze it, hoping to determine the underlying physical principles that govern the outcomes of our experiments.

Our ability to draw conclusions from sets of data rests on our computational power to run statistics. As computer computational power and accessibility continues to increase, more people have access to the tools and storage required to analyze large databases. Computationally intensive techniques that used to be impossible are now standard on home laptops.

Conclusion

When doing statistical analyses, or looking at others analyses, it is critical that we don't check our brains at the door. The application of statistical methods and the interpretation of results must involve both mathematics and logic, in order for the conclusions to be correct and the implications properly interpreted. We need to be clear about what a statistical statement does or does not imply. If we only consider formulas, we are likely to get incorrect or misleading results.

Statistics is more subtle than people may realize, so the underlying logic must be included. The choice of statistical summary, the method of data collection and/or sampling, bias in experimental construction, the arbitrary choices of thresholds for outliers or for invalidation of a hypothesis, all play a role in a result, and need to be considered when one is assessing the value and accuracy of that result.

In other words, statistics without context are meaningless.

At some level, statistics is just a quantification of ignorance. We can't make absolute statements, but we can assess the evidence to provide support for a hypothesis or conclusion, and also determine how confident we can be in our answer.