Lecture Notes 1
Econ 29000, Principles of Statistics
Kevin R Foster, CCNY
Fall 2011

The first part of this lecture covers "Know Your Data" and "Show Your Data," which reviews some of the very initial components necessary for data analysis.

You should view online video 1; that reviews basic information about measures of the data center such as mean, median, and mode; also measures of the spread of the data such as the standard deviation. Those notes are the middle part of this lecture. In class we will skip right to Lecture 2, where we apply these basic measures to learn about the ATUS dataset.

Further online material about statistics can be found:

- Hans Rosling is a phenom of TED talks and now "The Joy of Stats" here, http://www.open.ac.uk/openlearn/whats-on/the-joy-stats His website also has "The Joy of Stats," along with some data, http://www.gapminder.org/ His TED talks: http://www.ted.com/speakers/hans_rosling.html

- On Data Visualization, http://www.interaction-design.org/encyclopedia/

- Strata Conference (includes videos; Hilary Mason's is a good intro) http://strataconf.com/strata2011/public/schedule/proceedings

- http://www.scientificamerican.com/blog/post.cfm?id=words-pictures-and-the-visual-displ-2011-01-12

- Here's one on how lousy math education is: no real-world problem worth solving is set up like a textbook problem. Real-world problems are, well, problems – messy and incomplete. http://www.ted.com/talks/dan_meyer_math_curriculum_makeover.html His blog is http://blog.mrmeyer.com/ Explains why I give some of the homework assignments in such a format.

- Linear Regression "By Hand" from a *Wired* blog, http://www.wired.com/wiredscience/2011/01/linear-regression-by-hand/

- Using stats in unexpected ways: www.wired.com/magazine/2011/01/ff_lottery/all/1

- Vi Hart's blog is great for math stuff http://vihart.com/

These notes accompany the textbook used in the class, *Applied Statistics for Business and Economics*, David Doane and Lori Seward, 3rd edition, McGraw Hill.

If you begin a love affair with Statistics and want to read more, here are some suggestions:

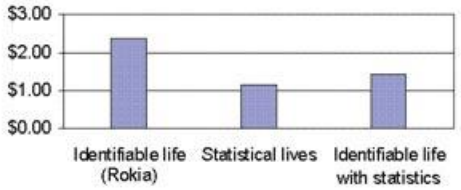
- Leonard Mlodinow, *Drunkard's Walk*
- Edward R. Tufte *The Visual Display of Quantitative Information*, *Visual Explanations: Images and Quantities, Evidence and Narrative* (in library)
- Howard Wainer, *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*
- David Salsburg, *Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*
- James Stock & Mark Watson, *Introduction to Econometrics* and Peter Kennedy, *A Guide to Econometrics*
- Jane E. Miller, *The Chicago Guide to Writing about Numbers* (in library)
- John W. Tukey, *Exploratory Data Analysis* (in library)
- Stephen Stigler, *Statistics on the Table* (in library) and *The History of Statistics: The Measurement of Uncertainty before 1900* (in library)
- Dierdre McCloskey , *Economical Writing* and *The Rhetoric of Economics* (in library)

The Challenge

Humans are bad at statistics, we're just not wired to think this way. Despite – or maybe, because of this, statistical thinking is enormously powerful and it can quickly take over your life. Once you begin thinking like a statistician you will begin to see statistical applications to even your most mundane activities.

Not only are humans bad at statistics but statistics seem to interfere with essential human feelings such as compassion.

> "A study by Small, Loewenstein, and Slovic (2007) ... gave people leaving a psychological experiment the opportunity to contribute up to $5 of their earnings to Save the Children. In one condition respondents were asked to donate money to feed an identified victim, a seven-year-old African girl named Rokia. They contributed more than twice the amount given by a second group asked to donate to the same organization working to save millions of Africans from hunger (see Figure 2). A third group was asked to donate to Rokia, but was also shown the larger statistical problem (millions in need) shown to the second group. Unfortunately, coupling the statistical realities with Rokia's story significantly reduced the contributions to Rokia.



> A follow-up experiment by Small et al. initially primed study participants either to feel ("Describe your feelings when you hear the word 'baby,'" and similar items) or to do simple arithmetic calculations. Priming analytic thinking (calculation) reduced donations to the identifiable victim (Rokia) relative to the feeling-based thinking prime. Yet the two primes had no distinct effect on statistical victims, which is symptomatic of the difficulty in generating feelings for such victims." *(Paul Slovic, Psychic Numbing and Genocide, November 2007, Psychological Science Agenda, http://www.apa.org/science/psa/slovic.html)*

Yet although we're not naturally good at statistics, it is very important for us to get better. Consider all of the people who play the lottery or go to a casino, sacrificing their hard-earned money. (Statistics questions are often best illustrated by gambling problems, in fact the science was pushed along by questions about card games and dice games.)

Google, one of the world's most highly-regarded companies, famously uses statistics to guide even its smallest decisions:

> A designer, Jamie Divine, had picked out a blue that everyone on his team liked. But a product manager tested a different color with users and found they were more likely to click on the toolbar if it was painted a greener shade.
>
> As trivial as color choices might seem, clicks are a key part of Google's revenue stream, and anything that enhances clicks means more money. Mr. Divine's team resisted the greener hue, so Ms. Mayer split the difference by choosing a shade halfway between those of the two camps.
>
> Her decision was diplomatic, but it also amounted to relying on her gut rather than research. Since then, she said, she has asked her team to test the 41 gradations between the competing blues to see which ones consumers might prefer (Laura M Holson, "Putting a Bolder Face on Google" New York Times, Feb 28, 2009).

Substantial benefits arise once you learn stats. Specifically, if so many people are bad at it then gaining a skill in Statistics gives you a scarce ability – and, since Adam Smith, economists have known that scarcity brings value. (And you might find it fun!)

Leonard Mlodinow, in his book *The Drunkard's Walk*, attributes the fact that we humans are bad at statistics as due to our need to feel in control of our lives. We don't like to acknowledge that so much of the world is genuinely random and uncontrollable, that many of our successes and failures might be due to chance. When statisticians watch games, we don't believe sportscasters who discuss "they just wanted it more" or other un-observable factors; we just believe that one team or the other got lucky.

As an example, suppose we were to have 1000 people toss coins in the air – those who get "heads" earn a dollar, and the game is repeated 10 times. It is likely that at least one person would flip "heads" all ten times. That person might start to believe, "Hey, I'm a good heads-tosser, I'm really good!" Somebody else is likely to have tossed "tails" ten times in a row – that person would probably be feeling stupid. But both are just lucky. And both have the same 50% chance of making "heads" on the next toss. Einstein famously said that he didn't like to believe that God played dice with the universe but many people look to the dice to see how God plays them.

Of course we struggle to exert control over our lives and hope that our particular choices can determine outcomes. But, as we begin to look at patterns of events due to many people's choices, then statistics become more powerful and more widely applicable. Consider a financial market: each individual trade may be the result of two people each analyzing the other's offers, trying to figure out how hard to press for a bargain, working through reams of data and making tons of calculations. But in aggregate, financial markets move randomly – if they did not then people could make a lot of money exploiting the patterns. Statistics help us both to see patterns in data that would otherwise see random and also to figure out when the patterns we observe are due to random chance. Statistics is an incredibly powerful tool.

Economics is a natural fit for statistical analysis since so much of our data is quantitative. Econometrics is the application of statistical analyses to economic problems. In the words of John Tukey, a legendary pioneer, we believe in the importance of "quantitative knowledge – a belief that most of the key questions in our world sooner or later demand answers to *by how much?* rather than merely to *in which direction?*"

This class

In my experience, too many statistics classes get off to a slow start because they build up gradually and systematically. That might not sound like a bad thing to you, but the problem is that you, the student, get answers to questions that you haven't yet asked. It can be more helpful to jump right in and then, as questions arise, to answer those at the appropriate time. So we'll spend a lot of time getting on the computer and actually doing statistics.

So the class will not always closely follow the textbook, particularly at the beginning. We will sometimes go in circles, first giving a simple answer but then returning to the most important questions for more study. The textbook proceeds gradually and systematically so you should read that to ensure that you've nailed down all of the details.

Statistics and econometrics are ultimately used for persuasion. First we want to persuade ourselves whether there is a relationship between some variables. Next we want to persuade other people whether there is such a relationship. Sometimes statistical theory can become quite Platonic in insisting that there is some ideal coefficient or relationship which can be discerned. In this class we will try to keep this sort of discussion to a minimum while keeping the "persuasion" rationale uppermost.

Step One: Know Your Data

The first step in any examination of data is to know that data – where did it come from? Who collected it? What is the sample of? What is being measured? Sometimes you'll find people who don't even know the units!

Then think about the units you actually want and make the necessary transformations. Typically we compare rates not just levels; we usually want to be able to ask a question like, "is that big or small?" Considering your classes, would you like to know how many A grades were earned in last year's class, or what fraction of the students got A grades? If you were told that 18 students got A grades, and you wanted to know if that were big or small, you would immediately have to ask, "how big was the class?" Was it a 300-person lecture hall or a 20-person seminar? The rate or fraction of A grades bundles up these two pieces of information into something that is more understandable.

Economists and business people often get figures in various units: levels, changes, percent changes (growth), log changes, annualized versions of each of those. We need to be careful and keep the differences all straight.

Annualized Data

At the simplest level, consider if some economic variable is reported to have changed by 100 in a particular quarter. As we make comparisons to previous changes, this is straightforward (was it more than 100 last quarter? Less?). But this has at least two possible meanings – only the footnotes or prior experience would tell the difference. It could imply that the actual change was 100, so if the item continued to change at that same rate throughout the year, it would change by 400 after 4 quarters. Or it could imply that the actual change was 25 and if the item continued to change at that same rate it would be 100 after 4 quarters – this is an annualized change. Most GDP figures are annualized. But you'd have to read the footnotes to make sure.

This distinction holds for growth rates as well. But annualizing growth rates is a bit more complicated than simply multiplying. (These are also distinct from year-on-year changes.)

CPI changes are usually reported as monthly changes (not annualized). GDP growth is usually annualized. So a 0.2% change in the month's CPI and a 2.4% growth in GDP are actually the same! Any data report released by a government statistical agency should carefully explain if any changes are annualized or "at an annual rate."

Seasonal adjustments are even more complicated, where growth rates might be reported as relative to previous averages. We won't yet get into that.

To annualize growth rates, we start from the original data (for now assume it's quarterly not monthly): suppose some economic series rose from 1000 in the first quarter to 1005 in the second quarter. This is a 0.5% growth from quarter to quarter (=0.005). To annualize that

growth rate, we ask what would be the total growth, if the series continued to grow at that same rate for four quarters.

This would imply that in the third quarter the level would be 1005*(1 + 0.005) =1005*(1.005) = 1000*(1.005)*(1.005) = 1000*$(1.005)^2$; in the fourth quarter the level would be 1000*(1.005) *(1.005)*(1.005) = 1000*$(1.005)^3$; and in the first quarter of next year the level would be 1000*(1.005) *(1.005) *(1.005) *(1.005) = 1000*$(1.005)^4$= 1020.2 which is a little more than 2%.

This would mean that the annualized rate of growth (for an item reported quarterly) would be the final value minus the beginning value, divided by the beginning value, which is

$$\frac{1000(1.005)^4 - 1000}{1000} = (1.005)^4 - 1.$$

Generalized, this means that quarterly growth is annualized by taking the single-quarter growth rate, g, and converting this to an annualized rate of $(1+g)^4 - 1$.

If this were monthly then the same sequence of logic would get us to insert a 12 instead of a 4 in the preceding formula. If the item is reported over t time periods, then the annualized rate is $(1+g)^t - 1$. (Daily rates could be calculated over 250 business days or 360 "banker's days" or 365/366 calendar days per year.)

The year-on-year growth rate is different. This looks back at the level from one year ago and finds the growth rate relative to that level.

Each method has its weaknesses. Annualizing needs the assumption that the growth could continue at that rate throughout the year – not always true (particularly in finance, where a stock could bounce by 1% in a day but it is unlikely to be up by over 250% in a year – there will be other large drops). Year-on-year changes can give a false impression of growth or decline after the change has stopped.

For example, if some item the first quarter of last year was 50, then it jumped to 60 in the second quarter, then stayed constant at 60 for the next two quarters, then the year-on-year change would be calculated as 20% growth even after the series had flattened.

Sometimes several measures are reported, so that interested readers can get the whole story. For examples, go to the US Economics & Statistics Administration, http://www.esa.doc.gov/, and read some of the "Indicators" that are released.

For example, on July 14, 2011, "The U.S. Census Bureau announced today that advance estimates of U.S. retail and food services sales for June, adjusted for seasonal variation and holiday and trading-day differences, but not for price changes, were $387.8 billion, an increase of 0.1 percent (±0.5%) from the previous month, and 8.1 percent (±0.7%)

above June 2010." That tells you the level (not annualized), the monthly (not annualized) growth, and the year-on-year growth. The reader is to make her own inferences.

GDP estimates are annualized, though, so we can read statements like this, from the BEA's July 29 release, "Current-dollar GDP ... increased 3.7 percent, or $136.0 billion, in the second quarter to a level of $15,003.8 billion. " The figure, $15 trillion, is scaled to an annual GDP figure; we wouldn't multiply by 4. On the other hand, the monthly retail sales figures above **are not** multiplied by 12.

So if, for instance, we wanted to know the fraction of GDP that is retail sales, we could **NOT** divide 387.8/15003.8 = 2.6%! Instead either multiply the retail sales figure by 12 **or** divide the GDP figure by 12. This would get 31%. More pertinently, if we hear that government stimulus spending added $20 billion, we might want to try to figure out how much this helped the economy. Again, dividing 20/15003.8 = 0.13% (13 bps) but this is wrong! The $15tn is at an annual rate but the $20bn is not, so we've got to get the units consistent. Either multiply 50 by 4 or divide 15,003.8 by 4. (This mistake has been made by even very smart people!)

So don't make those foolish mistakes and know your data. If you have a sample, know what the sample is taken from. Often we use government data and just casually assume that, since the producers are professionals, that it's exactly what I want. But "what I want" is not always "what is in the definition." Much government data (we'll be using some of it for this class) is based on the Current Population Survey (CPS), which represents the civilian non-institutional population. Since it's the main source of data on unemployment rates, it makes good sense to exclude people in the military (who have little choice about whether to go to work today) or in prison (again, little choice). But you might forget this, and wonder why there are so few soldiers in the data that you're working with <*forehead slap!*>.

So know your data. Even if you're using internal company numbers, you've got to know what's being counted – when are sales booked? Warehouse numbers aren't usually quite the same as accounting numbers.

Show the Data

One of the hottest fields currently is "Data Visualization." This arises from two basic facts: 1. We're drowning in data; and 2. Humans have good eyes.

We're drowning in data because increasing computing power makes so much more available to us. Companies can now consider giving top executives a "dashboard" where, just like a driver can tell how fast the car is travelling right now, the executive can see how much profit is being made right now. Retailers have automated scanners at the cash register and at the receiving bay doors; each store can figure out what's selling.

The data piles up while nobody's looking at it. An online store might generate data on the thousands of clicks simultaneously occurring, but it's probably just spooling onto some server's

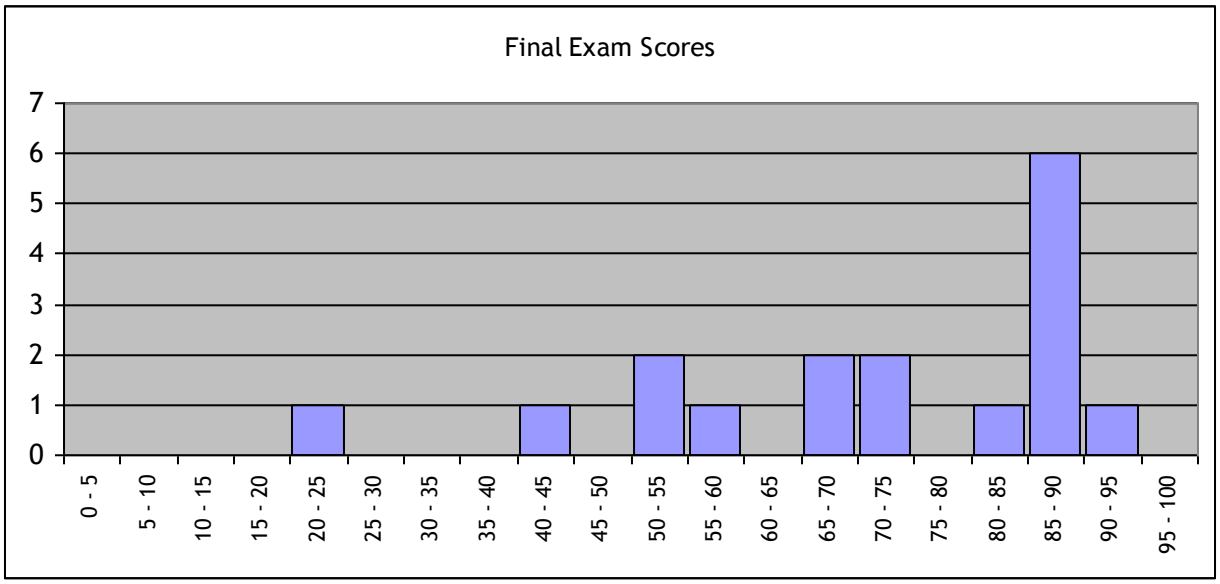
disk drive. It's just like spy agencies that harvest vast amounts of communications (voice, emails, videos, pictures) but then can't analyze them.

The hoped-for solution is to use our fundamental capacities to see patterns; convert machine data to visuals. Humans have good eyes; we evolved to live in the East African plains, watching all around ourselves to find prey or avoid danger. Modern people read a lot but that takes just a small fraction of the eye's nerves; the rest are peripheral vision. We want to make full use of our input devices.

But putting data into visual form is really tough to do well! The textbook has many examples to help you make better charts. Read Chapter 3 carefully. The homework will ask you to try your hand at it.

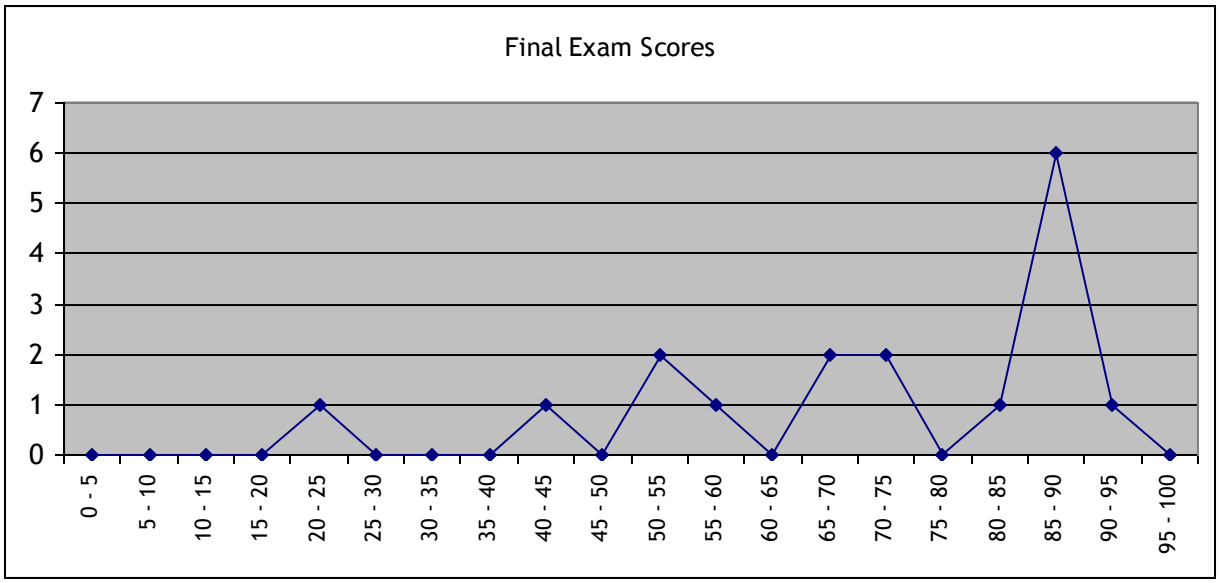
Histograms

You might have forgotten about histograms. A histogram shows the number (or fraction) of outcomes which fall into a particular bin. For example, here is a histogram of scores on the final exam for a class that I taught:



This histogram shows a great deal of information; more than just a single number could tell. (Although this histogram, with so many one- or two-step sizes, could be made much better.)

Often a histogram is presented, as above, with blocks but it can just as easily be connected lines, like this:

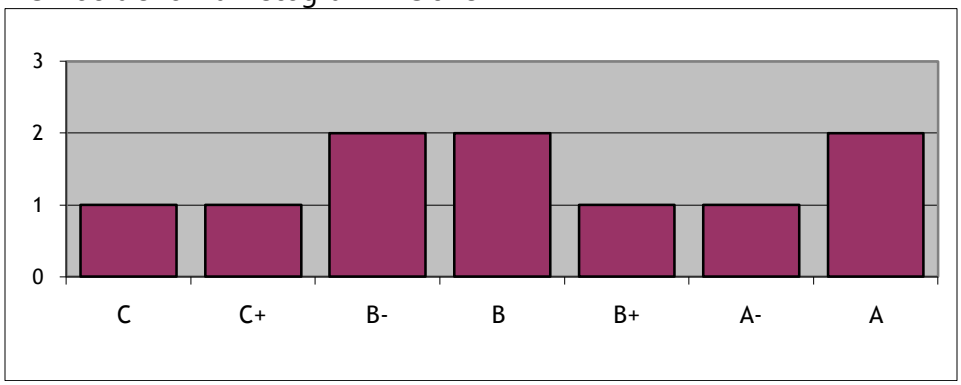


The information in the two charts is identical.

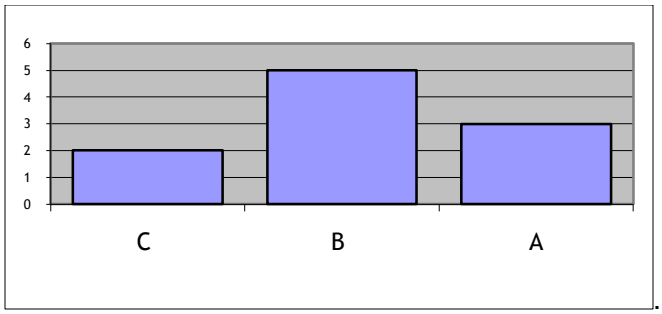
Histograms are a good way of showing how the data vary around the middle. This information about the spread of outcomes around the center is very important to most human decisions – we usually don't like risk.

Note that the choice of horizontal scaling or the number of bins can be fraught.

For example consider a histogram of a student's grades. If we leave in the A- and B+ grades, we would show a histogram like this:



whereas by collapsing together the grades into A, B, and C categories we would get something more intelligible, like this:



This shows the central tendency much better – the student has gotten many B grades and slightly more A grades than C grades. The previous histogram had too many categories so it was difficult to see a pattern.

The textbook has many more examples and suggestions as well as guides on what to avoid and how to improve your visual displays. Read Chapters 2 and 3 of Doane & Seward carefully.

Basic Concepts: Find the Center of the Data

You need to know how to calculate an average (mean), median, and mode. After that, we will move on to how to calculate measures of the spread of data around the middle, its variation.

Average

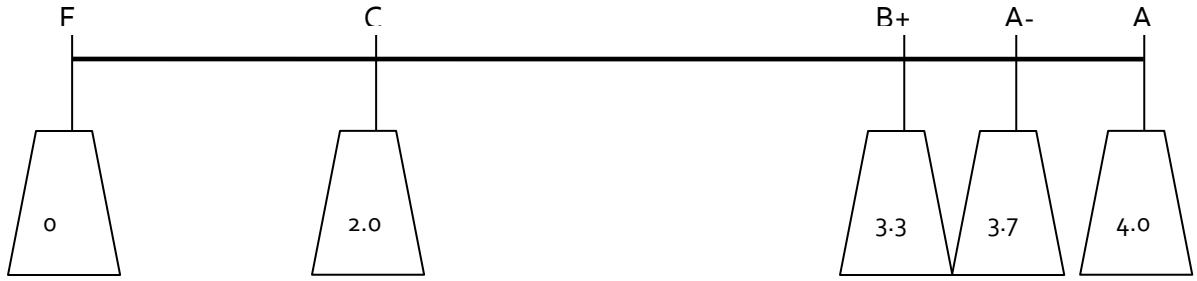
There are a few basic calculations that we start with. You need to be able to calculate an average, sometimes called the mean.

The average of some values, X, when there are N of them, is the sum of each of the values (index them by i) divided by N, so the average of X, sometimes denoted $\bar{X}$, is

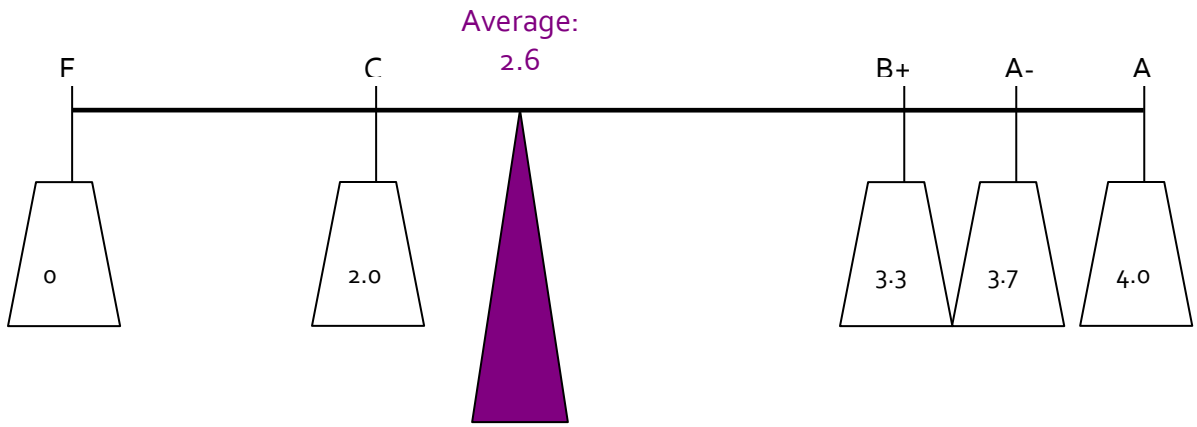
$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \, .$$

The average value of a sample is NOT NECESSARILY REPRESENTATIVE of what actually happens. There are many jokes about the average statistician who has 2.3 kids. If there are 100 employees at a company, one of whom gets a $100,000 bonus, then the average bonus was $1000 – but 99 out of 100 employees didn't get anything.

A common graphical interpretation of an average value is to interpret the values as lengths along which weights are hung on a see-saw. The average value is where a fulcrum would just balance the weights. Suppose a student is calculating her GPA. She has an A (worth 4.0), an A- (3.67), a B+ (3.33), a C (2.0) and one F (0) [she's having troubles!]. We could picture these as weights:



The weights "balance" at the average point (where (0 + 2 + 3.33 + 3.67 + 4)/5 = 2.6):



So the "bonus" example would look like this, with one person getting $100,000 while the other 99 get nothing:

Where there are actually 99 weights at "zero." But even one person with such a long moment arm can still shift the center of gravity away.

Bottom Line: The average is *often* a good way of understanding what happens to people within some group. But it is *not always* a good way.

Sometimes we calculate a weighted average using some set of weights, w, so

$$X_{weighted\ Average} = \sum_{i=1}^{n} w_i X_i \text{ , where } \sum_{i=1}^{n} w_i = 1.$$

Your GPA, for example, weights the grades by the credits in the course. Suppose you get a B grade (a 3.0 grade) in a 4-credit course and an A- grade (a 3.67 grade) in a 3-credit course; you'd calculate GPA by multiplying the grade times the credit, summing this, then dividing by the total credits:

$$GPA = \frac{3\cdot 4 + 3.67\cdot 3}{4+3} = \frac{4}{4+3}3 + \frac{3}{4+3}3.67 = 3.287.$$

So in this example the weights are $w_1 = \frac{4}{4+3}, w_2 = \frac{3}{4+3}$.

When an average is projected forward it is sometimes called the "Expected Value" where it is the average value of the predictions (where outcomes with a greater likelihood get greater weight). This nomenclature causes even more problems since, again, the "Expected Value" is NOT NECESSARILY REPRESENTATIVE of what actually happens.

To simplify some models of Climate Change, if there is a 10% chance of a 10° increase in temperature and a 90% chance of no change, then the calculated Expected Value is a 1° change – but, again, this value does not actually occur in any of the model forecasts.

For those of you who have taken calculus, you might find these formulas reminiscent of integrals – good for you! But we won't cover that now. But if you think of the integral as being just an extreme form o f a summation, so the formula has the same format.

Median

The median is another measure of what happens to a 'typical' person in a group; like the mean it has its limitations. The median is the value that occurs in the 50^{th} percentile, to the person (or occurrence) exactly in the middle. If there are an odd number of outcomes, otherwise it is between the two middle ones.

In the bonus example above, where one person out of 100 gets a $100,000 bonus, the median bonus is $0. The two statistics combined, that the average is $1000 but the median is zero, can provide a better understanding of what is happening. (Of course, in this very simple case, it is easiest to just say that one person got a big bonus and everyone else got nothing. But there may be other cases that aren't quite so extreme but still are skewed.)

Mode

The mode is the most common outcome; often there may be more than one. If there were a slightly more complicated payroll case, where 49 of the employees got zero bonus, 47 got $1000, and four got $13,250 each, the mean is the same at $1,000, the median is now equal to the mean [review those calculations for yourself!], but the mode is zero. So that gives us additional information beyond the mean or median.

Spread around the center

Data distributions differ not only in the location of their center but also in how much spread or variation there is around that center point. For example a new drug might promise an average of 25% better results than its competitor, but does this mean that 25% of patients improved by 100%, or does this mean that everybody got 25% better? It's not clear from just the central tendency. But if you're the one who's sick, you want to know.

This is a familiar concept in economics where we commonly assume that investors make a tradeoff between risk and return. Two hedge funds might both have a record of 10% returns, but a record of 9.5%, 10%, and 10.5% is very different from a record of 0%, 10%, and 20%. (Actually a record of always winning, no matter what, distinguished Bernie Madoff's fund…)

You might think to just take the average difference of how far observations are from the average, but this won't work.

There's an old joke about the tenant who complains to the super that in winter his apartment is 50° and in summer is 90° -- and the super responds, "Why are you complaining? The apartment is a comfortable 70° on average!" (So the tenant replies "*I'm complaining because I have a squared error loss function*!" If you thought that was funny, you're a stats geek already!)

The average deviation from the average is always zero. Write out the formulas to see.

The average of some N values, $X_1, X_2, \ldots X_N$, is given by $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$.

So what is the average deviation from the average, $\sum_{i=1}^{N}\left(X_i - \bar{X}\right)$?

We know that $\sum_{i=1}^{N}\left(X_i - \bar{X}\right) = \sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \bar{X}$ and, since $\bar{X}$ is the same for every observation, $\sum_{i=1}^{N} \bar{X} = N\bar{X} = \sum_{i=1}^{N} X_i$, if we substitute back from the definition of $\bar{X}$. So $\sum_{i=1}^{N}\left(X_i - \bar{X}\right) = 0$. We can't re-use the average. So we want to find some useful, sensible function [or functions], $f(\cdot)$, such that $\sum_{i=1}^{N} f\left(X_i - \bar{X}\right) \neq 0$.

Standard Deviation

The most commonly reported measure of spread around the center is the standard deviation. This looks complicated since it squares the deviations and then takes the square root, but is actually quite generally useful.

The formula for the standard deviation is a bit more complicated:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\,.$$

Before you start to panic, let's go through it slowly. First we want to see how far each observation is from the mean,

$$\left(X_i - \overline{X}\right).$$

If we were to just sum up these terms, we'd get nothing – the positive errors and negative errors would cancel out.

So we square the deviations and get

$$\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

and then just divide by n to find the average squared error, which is known as the variance, which is

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2\,.$$

The standard deviation is the square root of the variance; $\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$.

Of course you're asking why we bother to square all of the parts inside the summation, if we're only going to take the square root afterwards. It's worthwhile to understand the rationale since similar questions will re-occur. The point of the squared errors is that they don't cancel

out. The variance can be thought of as the average size of the squared distances from the mean. Then the square root makes this into sensible units.

The variance and standard deviation of the population divides by N; the variance and standard deviation of a sample divide by (N – 1). This is referred to as a "degrees of freedom correction," referring to the fact that a sample, after calculating the mean, has lost one "degree of freedom," so the standard deviation has only (N – df) remaining. You could worry about that difference or you could note that, for most datasets with huge N (like the ATUS with almost 100,000), the difference is too tiny to worry about.

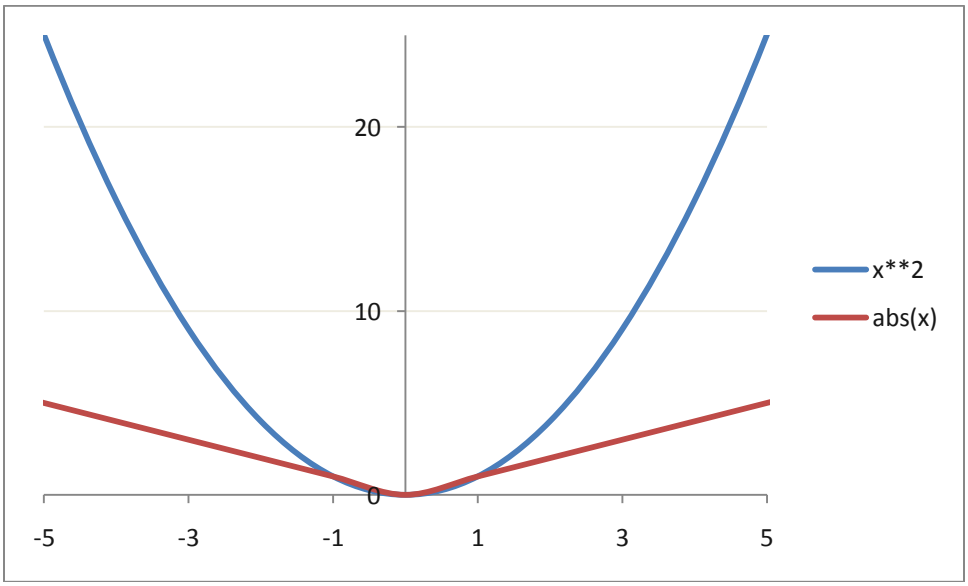
Our notation generally uses Greek letters to denote population values and English letters for sample values, so we have

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$ and

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$.

As you learn more statistics you will see that the standard deviation appears quite often. Hopefully you will begin to get used to it.

We could look at other functions of the distance of the data from the central measure, $f(\cdot)$, such that $\sum_{i=1}^{N} f\left(X_i - \bar{X}\right) \neq 0$ -- for example, the mean of the absolute value, $\frac{1}{N}\sum_{i=1}^{N}\left|X_i - \bar{X}\right|$. By recalling the graphs of these two functions you can begin to appreciate how they differ:



So that squaring the difference counts large deviations very much worse than small deviations, whereas an absolute deviation does not. So if you're trying to hit a central target, it might well make sense that wider and wider misses should be penalized worse, while tiny misses should be hardly counted.

There is a relationship between the distance measure selected and the central parameter. For example, suppose I want to find some number, Z, that minimizes a measure of distance of this number, Z, from each observations. So I want to minimize $\frac{1}{N}\sum_{i=1}^{N} f(X_i - Z)$. If we were to use the absolute value function then setting Z to the median would minimize the distance. If we use instead the squared function then setting Z to the average would minimize the distance. So there is an important connection between the average and the standard deviation, just as there is a connection between the median and the absolute deviation. *(Can you think of what distance measure is connected with the mode?)*

If you know calculus, you will understand why, in the age before computer calculations, statisticians preferred the squared difference to the absolute value of the difference. If we look for an estimator that will minimize that distance, then in general in order to minimize something we will take its derivative. But the derivative of the absolute value is undefined at zero, while the squared distance has a well-defined derivative.

Sometimes you will see other measures of variation; the textbook goes through these comprehensively. Note that the Coefficient of Variation, $\frac{s}{\overline{X}}$, is the reciprocal of the signal-to-noise ratio. This is an important measure when there is no natural or physical measure, for example a Likert scale. If you ask people to rate beers on a scale of 1-10 and find that consumers prefer Stone's Ruination Ale to Budweiser by 2 points, you have no idea whether 2 is a big or a small difference – unless you know how much variation there was in the data (i.e. the standard deviation). On the other hand, if Ruination costs $2 more than Bud, you can interpret that even without a standard deviation.

In finance, this signal/noise ratio is referred to as the Sharpe Ratio, $\frac{\overline{R} - r_f}{\sigma}$, where $\overline{R}$ are the average returns on a portfolio and r_f is the risk-free rate; the Sharpe Ratio tells the returns relative to risk.

Sometimes we will use "Standardized Data," usually denoted as Z_i, where the mean is subtracted and then we divide by the standard deviation, so $Z_i = \frac{X_i - \overline{X}}{s}$. This is interpretable as measuring how many standard deviations from the mean is any particular observation. This allows us to abstract from the particular units of the data (meters or feet; Celsius or Fahrenheit; whatever) and just think of them as generic numbers.

Now Do It!

We'll use data from the Census PUMS, on just people in New York City, to begin actually doing statistics using the analysis program called SPSS. There are further lecture notes on each of those topics. Read those carefully; you'll need them to do the homework assignment.

Next:

- on the PUMS data
- on using SPSS (also videos)
- Lecture Notes 2