Beginning Notes

Econ B2000, Statistics and Introduction to Econometrics

Kevin R Foster, the City College of New York, CUNY

Fall 2014

Preliminary

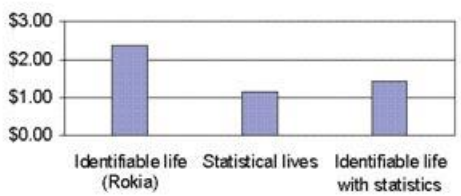
We begin with "Know Your Data" and "Show Your Data," to review some of the very initial components necessary for data analysis.

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 sports games, we don't believe sportscasters who discuss "that player 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!

Economists 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): 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*$(1.005)^4$, 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

A hot field 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.

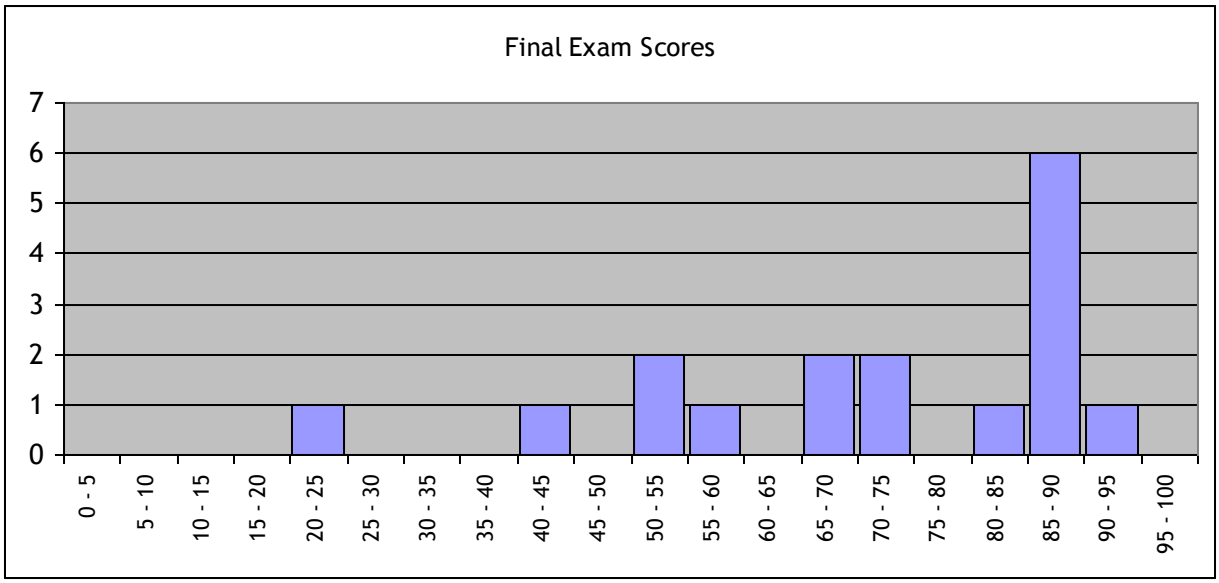
The data piles up while nobody's looking at it. An online store might generate data on the thousands of clicks simultaneously occurring, but it's probably just spooling onto some server's disk drive. It's just like spy agencies that harvest vast amounts of communications (voice, emails, videos, pictures) 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.

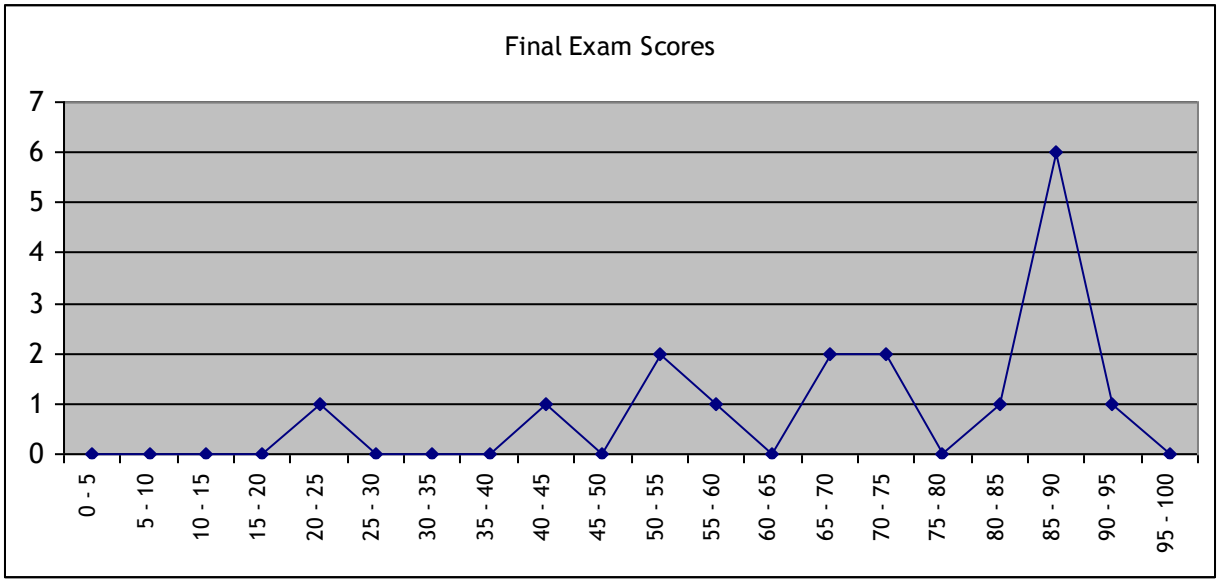
Histograms

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:

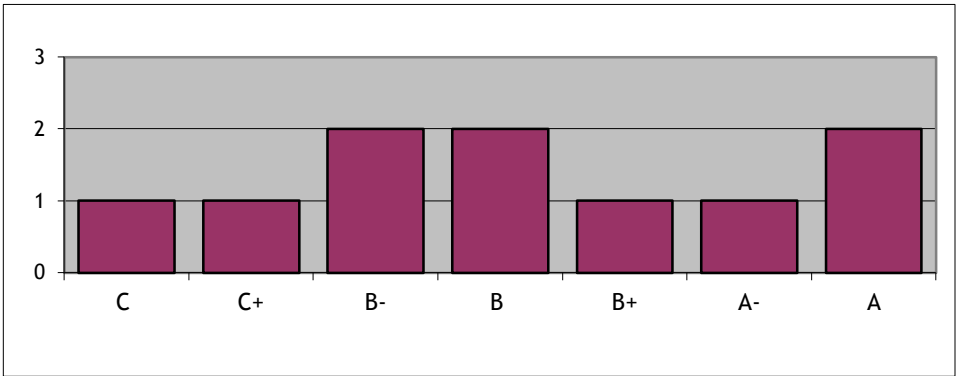


The information in the two charts is identical.

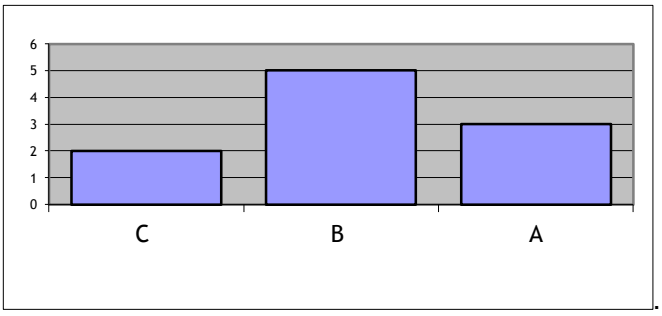
Histograms are a good way of showing how the data vary around the middle. This information about the spread of outcomes around the center is very important to most human decisions – we usually don't like risk.

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.

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.

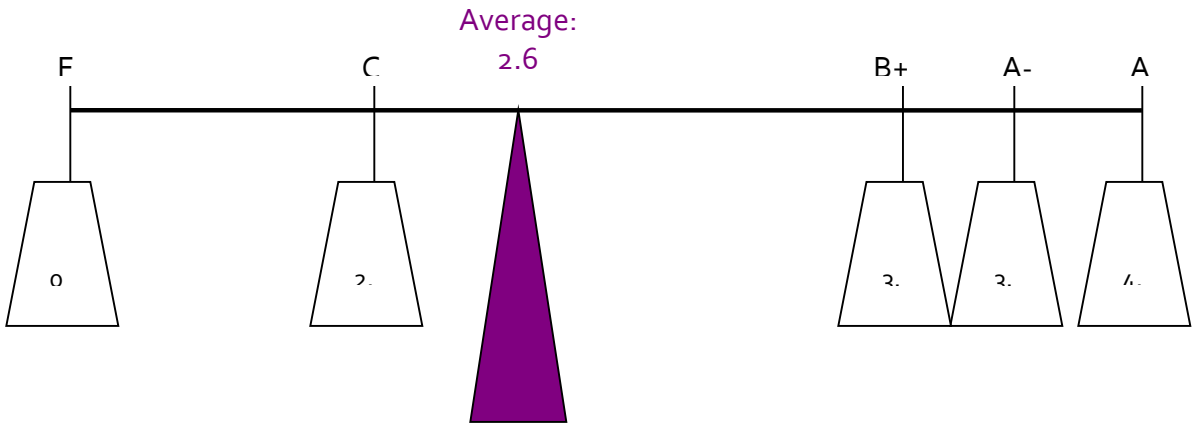
The average of some values, X, when there are N of them, is the sum of each of the values (index them by i) divided by N, so the average of X, sometimes denoted $\overline{X}$, is

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

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, then 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 50th 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

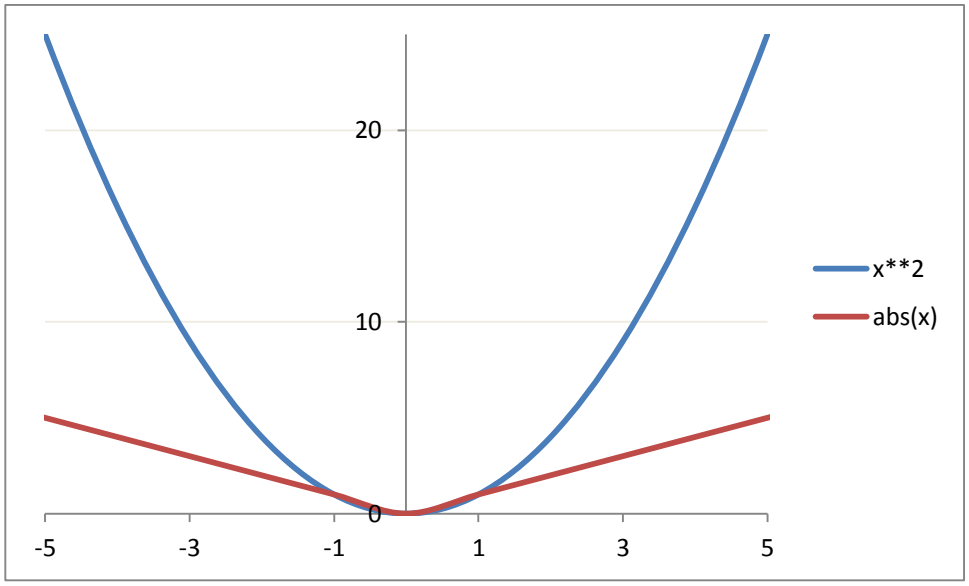
$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$ and

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

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, $\frac{s}{\bar{X}}$, 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 $Z_i = \frac{X_i - \bar{X}}{s}$. 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 R. There are further lecture notes on each of those topics. Read those carefully; you'll need them to do the homework assignment.

Overview of PUMS

We will use data from the Census Bureau's "Public Use Microdata Survey," or PUMS. This is collected in the American Community Survey; just about every ten years since 1790 the Census has made a complete enumeration of the US population as required by the Constitution. I got the data from IPUMS, which collects and makes available historical and contemporaneous Census data samples.

We will work on this data using R. Later I give an overview of the basics of how to use that program.

The dataset has information on 196,314 people in 85,730 households. If there is a family living together in an apartment, say a parent and two kids, then each person has a row of data telling about him/her (age, gender, education, etc) but only the head of household would have information about the household (how much is spent on rent, utilities, etc.). In this data, PERNUM gives the number of the person in the household; person #1 is the head of household (however they choose to answer). Depending on what analysis is to be made, the researcher might want to look at all the people or all of the households (or subsets of either). (Note that the "head of household" is defined by the person interviewed so it could be the man or woman, if there are both.)

There are variables coding people's race/ethnicity, if they were born in the US or a foreign country, how much schooling they have, if they are single or married, if they're a veteran, what borough they live in and how they commute to work. There is some greater detail about ancestry (where people can write in detail about their background). There is information about their incomes. For the household there is information about the dwelling including how much they spend on mortgage/rent, how many rooms, how many units, and when it was built.

About R

R is a popular and widely-used statistical program. It might seem a bit overwhelming at first but you will learn to appreciate it. Its main advantage is that it is open-source so there are many 'packages' built by statisticians to run particular specialized estimations. BTW that means it's free for you to download and install – which is surely another advantage. [1]

Why learn this particular program? You should not be monolingual in statistical analysis, it is always useful to learn more programs. The simplest is Excel, which is very widely used but has a number of limitations – mainly that, in order to make it easy for ordinary people to use, they made it tough for power users. SPSS is the next step: a bit more powerful but also a bit more difficult. Next are Stata and SAS. Matlab is great but proprietary so not as widely used. Python might be important in some careers. The college has R, SPSS, SAS, and Matlab freely available in all of the computer labs. R is powerful, versatile, and widely used. This site has detailed analytics of which software is most common in job posts and other measures, http://r4stats.com/articles/popularity/ .

You might be tempted to just use Excel; resist! Excel doesn't do many of the more complex statistical analyses that we'll be learning later in the course. Make the investment to learn a better program; trust me on the cost/benefit ratio.

I will give a basic (albeit short) overview in these notes. If you want a deeper review, I included the book, *A Beginner's Guide to R*, by Zuur, Ieno and Meesters as a recommended text for the course as well as *Applied Econometrics with R* by Kleiber and Zeileis.

The Absolute Beginning

Start up R. On any of the computers in the Economics lab (6/150) double-click on the "R" logo on the desktop to start up the program. In other computer labs you might have to do a bit more hunting to find R (if there's no link on the desktop, then click the "Start" button in the lower left-hand corner, and look at the list of "Programs" to find R).

If you're going to install it on your home computer, download the program from R-project.org (http://www.r-project.org/) and follow those instructions – it has versions for Mac, Windows, and various Unix builds. You might want to also download R Studio

[1] I am grateful to Herby Brutus who was test pilot for the first version of "About R" and made useful suggestions.

(http://www.rstudio.com/) which is a helper program that sits on top of R to make it a bit friendlier to work with. (That's one of the advantages of R being open-source – the original developers didn't much worry about friendliness, so somebody else did.) Of course if you have trouble, Google can usually help.

With either R or R-Studio, you'll get an old-fashioned command line called "Console" (just " > ") where you can type in (or copy-and-paste in) commands to the program.

Until you get used to it, this might seem like a cost not a benefit. But consider if you've ever dug through someone else's Excel sheet (or even an old one of your own), trying to figure out, "how did they ever get that number?!?" For the sake of simplification and ease of use, it loses replicability. It can be tough to replicate what someone else did – but replication is the basis of science. So a little program that shows what you did each time can actually be really important. It also means you can use other people's code and instructions.

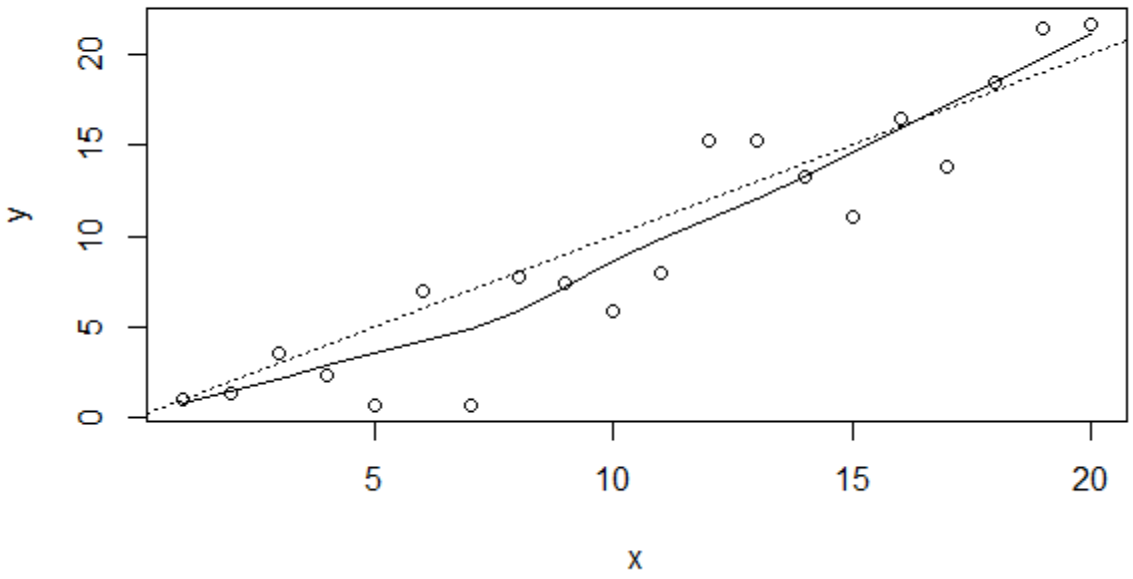
The guide, *An Introduction to R*, suggests these to give a basic flavor of what's done and hint at the power. Don't worry too much about each step for now, this is just a glimpse to show that with a few commands you can do some relatively sophisticated estimation – i.e. for a low cost you can get a large benefit.

```
x <- 1:20
w <- 1 + sqrt(x)/2
example1 <- data.frame(x=x, y= x + rnorm(x)*w)
attach(example1)
```

This creates x and y variables (where the `rnorm` command creates random numbers from a normal distribution), puts them into a data frame, then attaches that data frame so that R can use it in later calculations. Next some stats – create a linear model (that's the "lm") then a "lowess" nonparametric local regression, and plot the two estimations to compare. (Just copy and paste these, don't worry about understanding them for now!)

```
fm <- lm(y ~ x)
summary(fm)
lrf <- lowess(x, y)
plot(x, y)
lines(x, lrf$y)
abline(0, 1, lty=3)
abline(coef(fm))
detach()
```

You should get a graph looking something like this (although with the random numbers, not exactly):



The final "detach" command just cleans up, it is the opposite of "attach".

For all of these commands, you can use R to type "`help(__)`" where you fill in __ in the obvious way to get help on commands including how to make various changes. Or shortcut with just "`?__`" so for example type "`?summary`" or "`help(summary)`". But as I said, don't worry much about those commands for now, I'm not expecting you to become an R-ninja overnight.

Basics in R

We will start with a few commands to get you able to follow along with these notes. For now you will use data that I've put together for you, ready to use in R, beginning with the Census Bureau's PUMS (Public Use Microdata Sample, from the American Community Survey, accessed from IPUMS). Download that data from InYourClass, the file is `pums_NY.RData`. I have restricted the sample to contain only people living in the state of New York. You might want to move that into a new directory (at least remember what directory you put it in), maybe name the new folder/directory "`pums_NY`." The commands are in the file, "`working_on_PUMS.R`" so you might download that too.

Start with these commands in R or R-Studio; the first wipes the program clean:

```
rm(list = ls(all = TRUE))
setwd("C:\\pums_NY") # Change this as appropriate
load("pums_NY.RData")
```

The next, "setwd," is to set the working directory for this analysis. Your directory is somewhere on your computer, figure out the path name – in Windows it is usually something like `setwd("C:\\Users\\Kevin\\Documents\\R\\pums_NY")` and for Mac, `setwd("~/desktop/R/pums_NY")` – in both of those I'm assuming you created a folder called "R" and then inside that another folder, "pums_NY". The only difference is a doubled backslash in the name here in the program. As long as you had put the downloaded data into that directory and the program can find that directory, you should not have an error.

To check out the data use `str(dat_pums_NY)` which will show a list of the variables in the data and the first 10 or so lines of each variable, something like this

```
'data.frame':  196314 obs. of  57 variables:
 $ Age             : num  43 45 33 57 52 26 83 87 21 45 ...
 $ female          : num  1 0 0 0 1 0 1 0 0 0 ...
 $ PERNUM          : num  1 2 1 1 2 3 1 2 1 1 ...
```

(but longer!) In the next section I'll explain more about the data and what the lines mean but for now `Age` is the person's age in years and `female` is a logical 0/1 variable. So the first person is a 43-year-old female, next is a 45-year-old male, etc.

Next, run these

```
attach(dat_pums_NY)
NN_obs <- length(Age)
```

Which, as I said above, attaches the data frame to the program and then the next line should tell you how many people are in the data: 196,314.

Next we compare the average age of the men and the women in the data,

```
summary(Age[female == 1])
summary(Age[!female])
```

The female dummy variable is a logical term, with a zero or one for false or true. The comparison is between those who have the variable female=1 (i.e. women) and those not female=1 (so logical not, denoted with the "!" symbol, i.e. men). *I know, you can (and people do!) worry that this binary classification for gender misses some people; government statistics are just not there yet.* I find this output,

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   21.00   43.00   41.88   60.00   94.00

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   19.00   39.00   39.19   57.00   94.00
```

So women are, on average, a bit older, with an average age of 41.9 compared with 39.2 for men. You might wonder (if you were to begin to think like a statistician) whether that is a big difference – hold onto that thought!

Alternately you can use

```
mean(Age[female == 1])
sd(Age[female == 1])
mean(Age[!female])
sd(Age[!female])
```

to get mean and standard deviation of each. Later you might encounter cases where you want more complicated dummy variables and want to use logical relations "and" "or" "not" (the symbols "`&`", "`|`", "`!`") or the ">=" or multiplication or division.

As you're going along, if you're copying-and-pasting then you might not have had trouble, but if you're typing then you've probably realized that R is persnickety – the variable `Age` is not equivalent to a variable `AGE` nor `age` nor `aGe` ...

Rather than cutting-and-pasting all of these command lines from the internet or your favorite word processing software, you might want something a bit easier. R-Studio has a text editor or you can find one online (I like Notepad + +) to download and install. Don't use MS Word, the autocorrect will kill you. But if you save the list of commands as a single file, you can just run the whole list all at once – which is easier once you get into more sophisticated stuff. As I mentioned above, I've saved them into a file called "`working_on_PUMS.R`" and you can create your own. That's also a help for when you say, "D'oh!" and realize you have to go back and re-do some work – if it's all in a file then that's easy to fix. *(Feeling fancy, look into Sweave that combines the program with LATEX.)*

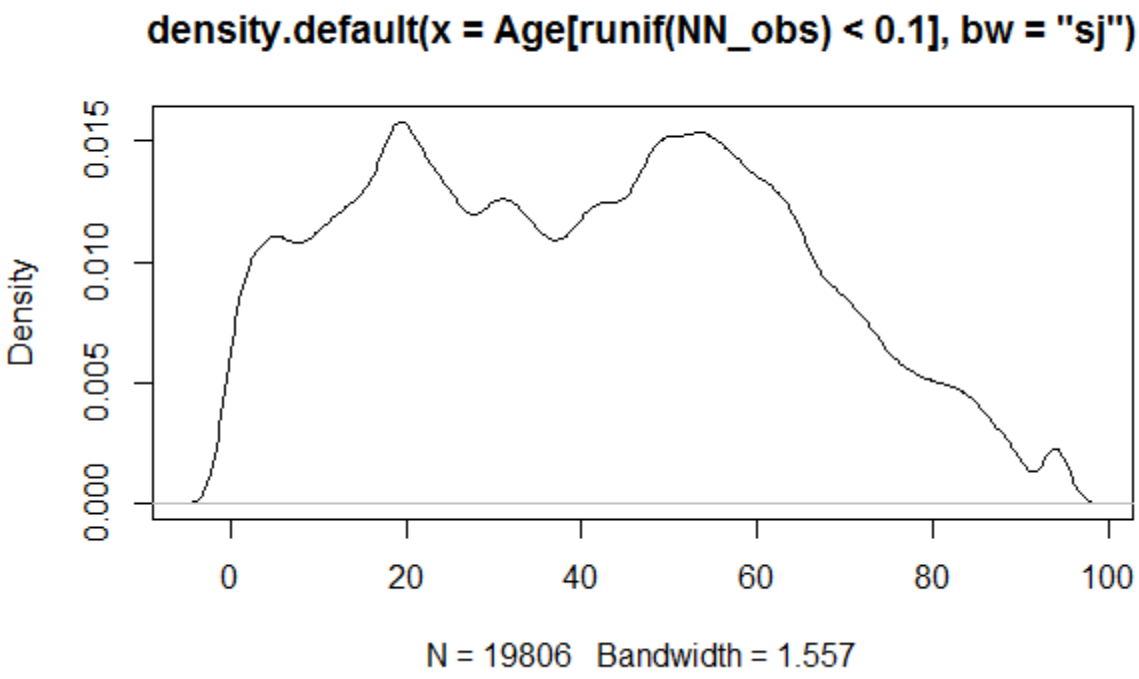
Before you get too far, remember to save your work. The computers in the lab wipe the memory clean when you log off so back up your data. Either online (email it to yourself or upload to Google Drive or iCloud or InYourClass) or use a USB drive.

Codes of Variables

Some of the PUMS variables here have a natural interpretation, for instance Age is measured in years. Actually even this has a bit of a twist. If you type "`hist(Age)`" you'll get a histogram of Age, which doesn't show any problem. But if you do a somewhat more sophisticated graph (a nonparametric kernel density),

```
plot(density(Age[runif(NN_obs) < 0.1], bw = "sj"))
```

you can get a fancier picture, like this:



Note, for you geeks, that if you tried this with the whole sample then the 200,000 observations would be too much, so I've randomly selected 10% (that's the `runif(NN_obs)<0.1` portion).

The kernel density is like a histogram with much smaller bins; in this case it shows a bit of weirdness particularly in the right, where it looks like there are suddenly a bunch of people who are 94 but nobody in between. This is due to a coding choice by the Census, where really old people are just labeled as "94". So if you were to get finicky (and every good statistician is!) you might go back to the calculations of averages previously and modify them all like this, `mean(Age[(female == 1)&(Age<90)])` to select just those who are female and who are coded as having age less than 90. *Also recall that this would not change the median values – which is one point in favor of that measure.* You go do that, I'll wait right here...

Anyway, I was saying that many variables have a natural measure, like Age measured in years (below 90 anyway). Others are logical variables (called dummies) like female, Hispanic, or married – there is a yes/no answer that is coded 1/0. *(Note that if you're creating these on your own it's good to give names that have that sort of yes/no answer, so a variable named 'female' is better than one named 'gender' where you'd have to remember who are 1 and who are 0.)*

Other variables, like PUMA, have no natural explanation at all, you have to go to the codebook (or, in this case, the file pums_initial_recoding.r) to find out that this is "Public Use Microdata Area" where 3801 codes for Washington Heights/Inwood, 3802 is Hamilton Heights/Manhattanville/West Harlem, etc. The program will happily calculate the average value for PUMA (type in `mean(PUMA)` and see for yourself!) but this is a meaningless value – the average neighborhood code value? If you want to select just people living in a particular neighborhood then you'd have to look at the list below.

PUMA	Neighborhood
3701	NYC-Bronx CD 8--Riverdale, Fieldston & Kingsbridge
3702	NYC-Bronx CD 12--Wakefield, Williamsbridge & Woodlawn
3703	NYC-Bronx CD 10--Co-op City, Pelham Bay & Schuylerville
3704	NYC-Bronx CD 11--Pelham Parkway, Morris Park & Laconia
3705	NYC-Bronx CD 3 & 6--Belmont, Crotona Park East & East Tremont
3706	NYC-Bronx CD 7--Bedford Park, Fordham North & Norwood
3707	NYC-Bronx CD 5--Morris Heights, Fordham South & Mount Hope
3708	NYC-Bronx CD 4--Concourse, Highbridge & Mount Eden
3709	NYC-Bronx CD 9--Castle Hill, Clason Point & Parkchester
3710	NYC-Bronx CD 1 & 2--Hunts Point, Longwood & Melrose
3801	NYC-Manhattan CD 12--Washington Heights, Inwood & Marble Hill
3802	NYC-Manhattan CD 9--Hamilton Heights, Manhattanville & West Harlem
3803	NYC-Manhattan CD 10--Central Harlem
3804	NYC-Manhattan CD 11--East Harlem
3805	NYC-Manhattan CD 8--Upper East Side
3806	NYC-Manhattan CD 7--Upper West Side & West Side
3807	NYC-Manhattan CD 4 & 5--Chelsea, Clinton & Midtown Business District
3808	NYC-Manhattan CD 6--Murray Hill, Gramercy & Stuyvesant Town
3809	NYC-Manhattan CD 3--Chinatown & Lower East Side
3810	NYC-Manhattan CD 1 & 2--Battery Park City, Greenwich Village & Soho
3901	NYC-Staten Island CD 3--Tottenville, Great Kills & Annadale
3902	NYC-Staten Island CD 2--New Springville & South Beach
3903	NYC-Staten Island CD 1--Port Richmond, Stapleton & Mariner's Harbor
4001	NYC-Brooklyn CD 1--Greenpoint & Williamsburg
4002	NYC-Brooklyn CD 4—Bushwick
4003	NYC-Brooklyn CD 3--Bedford-Stuyvesant
4004	NYC-Brooklyn CD 2--Brooklyn Heights & Fort Greene
4005	NYC-Brooklyn CD 6--Park Slope, Carroll Gardens & Red Hook
4006	NYC-Brooklyn CD 8--Crown Heights North & Prospect Heights
4007	NYC-Brooklyn CD 16--Brownsville & Ocean Hill
4008	NYC-Brooklyn CD 5--East New York & Starrett City
4009	NYC-Brooklyn CD 18--Canarsie & Flatlands
4010	NYC-Brooklyn CD 17--East Flatbush, Farragut & Rugby
4011	NYC-Brooklyn CD 9--Crown Heights South, Prospect Lefferts & Wingate
4012	NYC-Brooklyn CD 7--Sunset Park & Windsor Terrace
4013	NYC-Brooklyn CD 10--Bay Ridge & Dyker Heights
4014	NYC-Brooklyn CD 12--Borough Park, Kensington & Ocean Parkway
4015	NYC-Brooklyn CD 14--Flatbush & Midwood
4016	NYC-Brooklyn CD 15--Sheepshead Bay, Gerritsen Beach & Homecrest
4017	NYC-Brooklyn CD 11--Bensonhurst & Bath Beach
4018	NYC-Brooklyn CD 13--Brighton Beach & Coney Island

4101 NYC-Queens CD 1--Astoria & Long Island City
4102 NYC-Queens CD 3--Jackson Heights & North Corona
4103 NYC-Queens CD 7--Flushing, Murray Hill & Whitestone
4104 NYC-Queens CD 11--Bayside, Douglaston & Little Neck
4105 NYC-Queens CD 13--Queens Village, Cambria Heights & Rosedale
4106 NYC-Queens CD 8--Briarwood, Fresh Meadows & Hillcrest
4107 NYC-Queens CD 4--Elmhurst & South Corona
4108 NYC-Queens CD 6--Forest Hills & Rego Park
4109 NYC-Queens CD 2--Sunnyside & Woodside
4110 NYC-Queens CD 5--Ridgewood, Glendale & Middle Village
4111 NYC-Queens CD 9--Richmond Hill & Woodhaven
4112 NYC-Queens CD 12--Jamaica, Hollis & St. Albans
4113 NYC-Queens CD 10--Howard Beach & Ozone Park
4114 NYC-Queens CD 14--Far Rockaway, Breezy Point & Broad Channel

You could find the average age of women/men living in a particular neighborhood by adding in another "&" to the previous line so `mean(Age[(female == 1)&(Age<94)&(PUMA == 4102)])` for Jackson Heights. (Or find the average age of each PUMA with `tapply(Age, PUMA, mean)` which, I admit, is a rather ugly bit of code.)

If you do a lot of analysis on a particular subgroup, it might be worthwhile to create a subset of that group, so that you don't have to always add on logical conditions. This can be done with the expressions:

```
restrict1 <- as.logical(Age >= 25)
dat_age_gt_25 <- subset(dat_pums_NY, restrict1)
```

So then you `detach()` the original dataset and instead `attach(dat_age_gt_25)`. Then any subsequent analysis would be just done on that subset. Just remember that you've done this (again, this is a good reason to save the commands in a program) otherwise you'll wonder why you suddenly don't have any kids in the sample.

You might be tired and bored by these details, but note that there are actually important choices to be made here, even in simply defining variables. Take the fraught American category of "race". This data has a variable, RACED, showing how people chose to classify themselves, as 'White,' 'Black,' 'American Indian or Alaska Native,' (plus enormous detail of which tribe), Asian, various combinations, and many more codes.

Suppose you wanted to find out how many Asians are in a particular population. You could count how many people identify themselves as Asian only; you could count how many people identify as Asian in any combination. Sometimes the choice is irrelevant; sometimes it can skew the final results (e.g. the question in some areas, are there more blacks or more Hispanics?).

Again, there's no "right" way to do it because there's no science in this peculiar-but-popular concept of "race". People's conceptions of themselves are fuzzy and complicated; these measures are approximations.

Basics of government race/ethnicity classification

The US government asks questions about people's race and ethnicity. These categories are social constructs, which is a fancy way of pointing out that they are not based on hard science but on people's own views of themselves (influenced by how people think that other people think of them...). Currently the standard classification asks people separately about their "race" and "ethnicity" where people can pick labels from each category in any combination.

The "race" categories include: "White alone," "Black or African-American alone," "American Indian alone," "Alaska Native alone," "American Indian and Alaska Native tribes specified; or American Indian or Alaska native, not specified and no other race," "Asian alone," "Native Hawaiian and other Pacific Islander alone," "Some other race alone," or "Two or more major race groups." Then the supplemental race categories offer more detail.

These are a peculiar combination of very general (well over 40% of the world's population is "Asian") and very specific ("Alaska Native alone") representing a peculiar history of popular attitudes in the US. Only in the 2000 Census did they start to classify people in mixed races. If you were to go back to historical US Censuses from more than a century ago, you would find that the category "race" included separate entries for Irish and French and various other nationalities. Stephen J Gould has a fascinating book, *The Mismeasure of Man*, discussing how early scientific classifications of humans tried to "prove" which nationalities/races/groups were the smartest. Ta-Nehisi Coates says, "racism invented race in America."

Note that "Hispanic" is not "race" but rather ethnicity (includes various other labels such as Spanish, Latino, etc.). So a respondent could choose "Hispanic" and any race category – some choose "White," some choose "Black," some might be combined with any other of those complicated racial categories.

If you wanted to create a variable for those who report themselves as African-American and Hispanic, you'd use the expression `(AfAm == 1) & (Hispanic == 1)`; sometimes stats report for non-Hispanic whites so `(white == 1) & (Hispanic != 1)`. You can create your own classifications depending on what questions you're investigating.

The Census Bureau gives more information here, http://www.census.gov/newsroom/minority_links/minority_links.html

All of these racial categories make some people uneasy: is the government encouraging racism by recognizing these classifications? Some other governments choose not to collect race data. But that doesn't mean that there are no differences, only that the government

doesn't choose to measure any of these differences. In the US, government agencies such as the Census and BLS don't generally collect data on religion.

Re-Coding complicated variables from initial data

If we want more combinations of variables then we create those. Usually a statistical analysis spends a lot of time doing this sort of housekeeping – dull but necessary.

Educational attainment is also classified with complicated codes: the original data has code 63 to mean high school diploma, 64 for a GED, 65 for less than a year of college, etc. I have transformed them into a series of dummy variables, zero/one variables for whether a person has no high school diploma, just a high school diploma, some college (but no degree), an associate's degree, a bachelor's degree, or an advanced degree. An advantage of these is that finding the mean of a zero/one variable gives the fraction of the sample who have a one. So if we wanted to find how many adults have various educational qualifications, we can use mean(educ_nohs[Age >= 25]) etc, to find that 14% have no high school, 28% have just a high school diploma, 16% have some college but no degree, 9% have an associate's, 18% have a bachelor's, and 15% have an advanced degree. You could do this by borough or neighborhood to figure out which places have the most/least educated people. *(Ahem! You <cough> COULD do that right now. Or wait for the homework assignment.)*

You can look at the variables in this dataset by the simple command `str(dat_pums_NY)` which gives the name of each variable along with the first few data points of each. (I did this a few sections ago, so this is review.) For this data it will show

```
 $ Age              : num  43 45 33 57 52 26 83 87 21 45 ...
 $ female           : num  1 0 0 0 1 0 1 0 0 0 ...
 $ PERNUM           : num  1 2 1 1 2 3 1 2 1 1 ...
```

so the first person in the data is aged 43 and is female. PERNUM is the Person Number in the household, so each new value of 1 indicates a new household. So the first household has the 43-year-old female and a 45-year-old male, then the second household has just a 33-y-o male, etc. You should look over the other variables in the data; the end of the file has some of the codes – for example the Ancestry 1 and 2 variables have enormously detailed codings of how people state their ancestry.

How to install packages

R depends crucially on "packages" – that's the whole reason that the open-source works. Some statistician invents a cool new technique, then writes up the code in R and makes it available. If you used a commercial package you'd have to wait a decade for them to update it; in R it's here now. Also if somebody hacks a nicer or easier way to do stuff, they write it up.

Some of the programs I use will depend on R packages. Installing them is a 2-step process: first, from R Studio, choose "Tools \ Install Packages" from the menu (from plain R, it's "Packages \ Install Packages"). Then in the command line or program, type `library(packagename)`. (Fill in name of package for "packagename".)

Time Series in R

This next part is mostly optional, just again showing some of the stuff that R can easily do for time series data. First install some packages,

```
library(zoo)
library(lattice)
library(latticeExtra)
library(gdata)
rm(list = ls(all = TRUE))
```

Then get data from online – also a cool easy thing to do with R.

```
# original data from:
oilspot_url <-
"http://www.eia.gov/dnav/pet/xls/PET_PRI_SPT_S1_D.xls"
oilspot_dat <- read.xls(oilspot_url, sheet = 2, pattern =
"Cushing")

oilfut_url <-
"http://www.eia.gov/dnav/pet/xls/PET_PRI_FUT_S1_D.xls"
oilfut_dat <- read.xls(oilfut_url, sheet = 2, pattern =
"Cushing, OK Crude Oil Future Contract 1")
```

Use R's built-in system for converting jumbles of letters and numbers into dates,

```
date_spot <- as.Date(oilspot_dat$Date, format='%b %d%Y')
date_fut <- as.Date(oilfut_dat$Date, format='%b %d%Y')
```

Then R's "ts" for time-series and the "zoo" package.

```
wti_spot <-
ts(oilspot_dat$Cushing..OK.WTI.Spot.Price.FOB..Dollars.per.Barre
l., start = c(1986,2), frequency = 365)
wti_fut1 <-
ts(oilfut_dat$Cushing..OK.Crude.Oil.Future.Contract.1..Dollars.p
er.Barrel., start = c(1983,89), frequency = 365)

wti_sp_dat <- zoo(wti_spot,date_spot)
wti_ft_dat <- zoo(wti_fut1,date_fut)

wti_spotfut <- merge(wti_sp_dat,wti_ft_dat, all=FALSE)
```

And plot the results.

```
plot(wti_spotfut, plot.type = "single", col = c("black",
"blue"))

# tough to see any difference, so try this
wti_2013 <- window(wti_spotfut, start = as.Date("2013-01-
01"), end = as.Date("2013-12-31"))
```

```
     plot(wti_2013, plot.type = "single", col = c("black",
"blue"))

     # if you like this publication, you can get fancier...
     asTheEconomist(xyplot(wti_2013, xlab="Cushing WTI Spot
Future Price",))
```

De-bugging

Without a doubt, programming is tough. In R or with any other program, it is frustrating and complicated and difficult to do it the first few times. Some days it feels like a continuous battle just to do the simplest thing! Keep going despite that, keep working on it.

Your study group will be very helpful of course.

There are lots of online resources for learning R; like *R for Beginners* by Paradis, or the main intro from the R website, *An Introduction to R*.

I mentioned some books at the beginning, *A Beginner's Guide to R*, by Zuur, Ieno and Meesters and *Applied Econometrics with R* by Kleiber and Zeileis.

Then there are all of the websites, including:

- http://flowingdata.com/2012/06/04/resources-for-getting-started-with-r/
- http://statmaster.sdu.dk/bent/courses/ST501-2011/Rcard.pdf with lists of common commands
- http://www.cookbook-r.com/ a "cookbook" for R
- http://www.statmethods.net/ with "Quick R"
- http://www.r-bloggers.com/

If you have troubles that you can't solve, email me for help. But try to narrow down your question: if you run 20 lines of code that produce an error, is there a way to reproduce the error after just 5 lines? What if you did the same command on much simpler data, would it still cause an error? Sending emails like "I have a problem with errors" might be cathartic but is not actually useful to anyone. If you've isolated the error and read the help documentation on that command, then you're on your way to solving the problem on your own.

Other Datasets

The class will use a number of other data sets, which I will provide to you already formatted for R. These are usually assembled by government bureaucrats who love their acronyms so they include names like Fed SCF, NHIS, BRFSS, NHANES, WVS, PUMS.

Overview of ATUS data

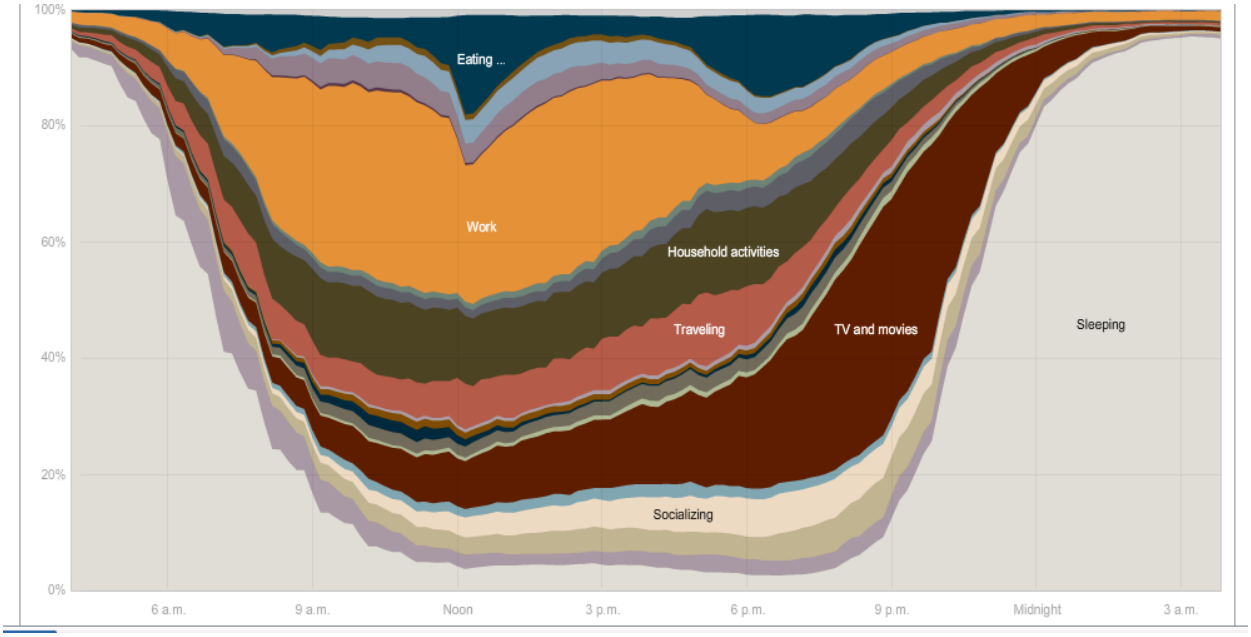
We will also use data from the "American Time Use Survey," or ATUS. This asks respondents to carefully list how they spent each hour of their time during the day; it's a tremendous resource. The survey data is collected by the US Bureau of Labor Statistics (BLS), a US government agency. You can find more information about it here, http://www.bls.gov/tus/.

The dataset has information on ## people interviewed from 2003-2013. This gives you a **ton** of information – we really need to work to get even the simplest information from it.

The dataset is ready to use in R. The ATUS has data telling how many minutes each person spent on various activities during the day. These are created from detailed logbooks that each person kept, recording their activities throughout the day.

They recorded how much time was spent with family members, with spouse, sleeping, watching TV, doing household chores, working, commuting, going to church/religious ceremonies, volunteering – there are hundreds of specific data items!

The NY Times had this graphic showing the different uses of time during the day [here http://www.nytimes.com/interactive/2009/07/31/business/20080801-metrics-graphic.html is the full interactive chart where you can compare the time use patterns of men and women, employed and unemployed, and other groups – a great way to lose an evening! The article is here http://www.nytimes.com/2009/08/02/business/02metrics.html?_r=2]



To use the data effectively, it is helpful to understand the ATUS classification system, where additional numbers at the right indicated additional specificity. The first two digits give generic broad categories. The general classification **T05** refers to time spent doing things

related to work. **T0501** is specific to actual work; **T050101** is "Work, main job" then **T050102** is "Work, other job," **T050103** is "Security Procedures related to work," and **T050189** is "Working, Not Elsewhere Classified," abbreviated as n.e.c. (usually if the final digit is a nine then that means that it is a miscellaneous or catch-all category). Then there are activities that are strongly related to work, that a person might not do if they were not working at a particular job – like taking a client out to dinner or golfing. These get their own classification codes, **T050201**, **T050202**, **T050203**, **T050204**, or **T050289**. The list continues; there are "Income-generating hobbies, crafts, and food" and "Job interviewing" and "Job search activities." These have other classifications beginning with **T05** to indicate that they are work-related.

So for instance, to create a variable, "Time Spent Working" that we might label "T_work," you would add up T050101, T050102, T050103, T050189, T050201, T050202, T050203, T050204, T050289, T050301, T050302, T050303, T050304, T050389, T050403, T050404, T050405, T050481, T050499, and T059999. You might want to add in "Travel related to working" down in T180501. (No sane human would remember all these codings but you'd look at the "Labels" and create a new variable.) It's tedious but not difficult in any way.

Some variables are even more detailed – playing sports is broken down into aerobics, baseball, basketball, biking, billiards, boating, bowling, ... all the way to wrestling, yoga, and "Not Elsewhere Classified" for those with really obscure interests. Then there are similar breakdowns for watching those sports. Most people will have a zero value for most of these but they're important for a few people.

You can imagine that different researchers, exploring different questions, could want different aggregates. So the basic data has a very fine classification which you can add up however you want.

Fed SCF, Survey of Consumer Finances produced by the Federal Reserve

This survey is only made once every three years; the most recent data is from 2010. The survey gives a tremendous amount of information about people's finances: how much they have in bank accounts (and how many bank accounts), credit cards, mortgages, student loans, auto and other loans, retirement savings, mutual funds, other assets – the whole panoply of financial information. But there's a catch. As you probably know from class as well as from personal experience, wealth is very unequally distributed. Some people have few financial assets at all, not even a bank account. Many people have only a few basic financial instruments: a credit card, some basic loans and a simple bank account. Then a few wealthy people have tremendously complicated portfolios of assets.

How does a statistical survey deal with this? By unequal sampling then weighting – all of the samples I provide here do this to one degree or another, but it becomes very important in the Fed SCF. The idea is simple: from the perspective of a survey about finance, all people with no financial assets look the same – they have "zero" for most answers in the survey. So a single response is an accurate sample for lots and lots of people. But people with lots of financial assets have varied portfolios, so a single response is an accurate sample for only a

small number of people. So if I were tasked with finding out about the financial system but could only survey 10 people, I might reasonably choose to sample 8 rich people with complicated portfolios and maybe 1 middle-class person and 1 poor person. I would keep in mind that the population of people in the country are not 80% rich, of course! In somewhat fancier statistics, I would weight each person, so the poor person would represent tens of millions of Americans, the middle-class person might represent more than a hundred million, and the rich people would each only represent a few million. If I wanted to extrapolate from the sample to the population, I would have to use these weights.

Many of the surveys we'll be using in class are weighted, and if you want to use them correctly you'll have to do the weighted versions. I'm skipping that for this class only because I think the cost outweighs the benefits for students early in their curriculum.

Actually using the Fed SCF survey can be difficult because the information is so richly detailed. You might want, say, a family's total debt, but instead get debt on credit card #1, card #2, all types of different loans, etc. so you have to add them up yourself. You have to do a bit of preliminary work.

NHIS National Health Interview Survey

This dataset has all sorts of medical and healthcare data – who has insurance, how often they're sick, doctor visits, pregnancy, weight/height. In the US many people have health insurance provided through their work so the economics of health and economics of insurance become tangled together.

BRFSS, Behavioral Risk Factor Surveillance System Survey

This dataset has many observations on a wide variety of risky behaviors: smoking, drinking, poor eating, flu shots, whether household has a 3-day supply of food and water... There is some economic data such as a person's income group.

NHANES – National Health And Nutrition Examination Survey

This has even more detail but on a smaller sample than the BRFSS. On whether people have healthy lifestyles: eat veg and fruit, their BMI, whether they smoke (various things), use drugs, sex (number of partners) – lots of things that are interesting enough to compensate for the dull (!?!?) stats necessary to analyze it.

There are other common data sources that are easily available online, which you can consider as you reflect upon your final project.

IPUMS

This is a tremendous data source, that has historical census data for past centuries, from http://www.ipums.org/. Some of the historical questions are weird (they asked if a person

was "idiotic" or "dumb" – which sounds crazy but used to be scientific terms). It includes full names and addresses from long-ago census data.

WVS World Values Survey

This has a bit less economics but still lots of interesting survey data about attitudes of people of many issues; the respondents are global from scores of countries over several different years. There is some information about personal income, education and occupation so you can see how those correlate with, say, attitudes toward democracy, religiosity, or other hot issues.

Demographic and Health Surveys from USAID

These give careful data about people in developing countries, to look at, say, how economic growth impacts nourishment.

Consumer Expenditure Data

Tons of data about household consumption patterns: how much they spend on shelter, transportation, food, gadgets, etc.

On Correlations: Finding Relationships between Two Variables

In a case where we have two variables, X and Y, we want to know how or if they are related, so we use covariance and correlation.

Suppose we have a simple case where X has a two-part distribution that depends on another variable, Y, where Y is what we call a "dummy" variable: it is either a one or a zero but cannot have any other value. (Dummy variables are often used to encode answers to simple "Yes/No" questions where a "Yes" is indicated with a value of one and a "No" corresponds with a zero. Dummy variables are sometimes called "binary" or "logical" variables.) X might have a different mean depending on the value of Y.

There are millions of examples of different means between two groups. GPA might be considered, with the mean depending on whether the student is a grad or undergrad. Or income might be the variable studied, which changes depending on whether a person has a college degree or not. You might object: but there are lots of other reasons why GPA or income could change, not just those two little reasons – of course! We're not ruling out any further complications; we're just working through one at a time.

In the PUMS data, X might be "wage and salary income in past 12 months" and Y would be male or female. Would you expect that the mean of X for men is greater or less than the mean of X for women?

Run this on R …

In a case where X has two distinct distributions depending on whether the dummy variable, Y, is zero or one, we might find the sample average for each part, so calculate the average when Y is equal to one and the average when Y is zero, which we denote $\left(\bar{X}\,|Y=0\right), \left(\bar{X}\,|Y=1\right)$ *or* $\bar{X}_{Y=0}, \bar{X}_{Y=1}$. These are called conditional means since they give the mean, conditional on some value.

In this case the value of $\bar{X}\,|Y=1$ is the same as the average of the two variables multiplied together, $X \cdot Y$.

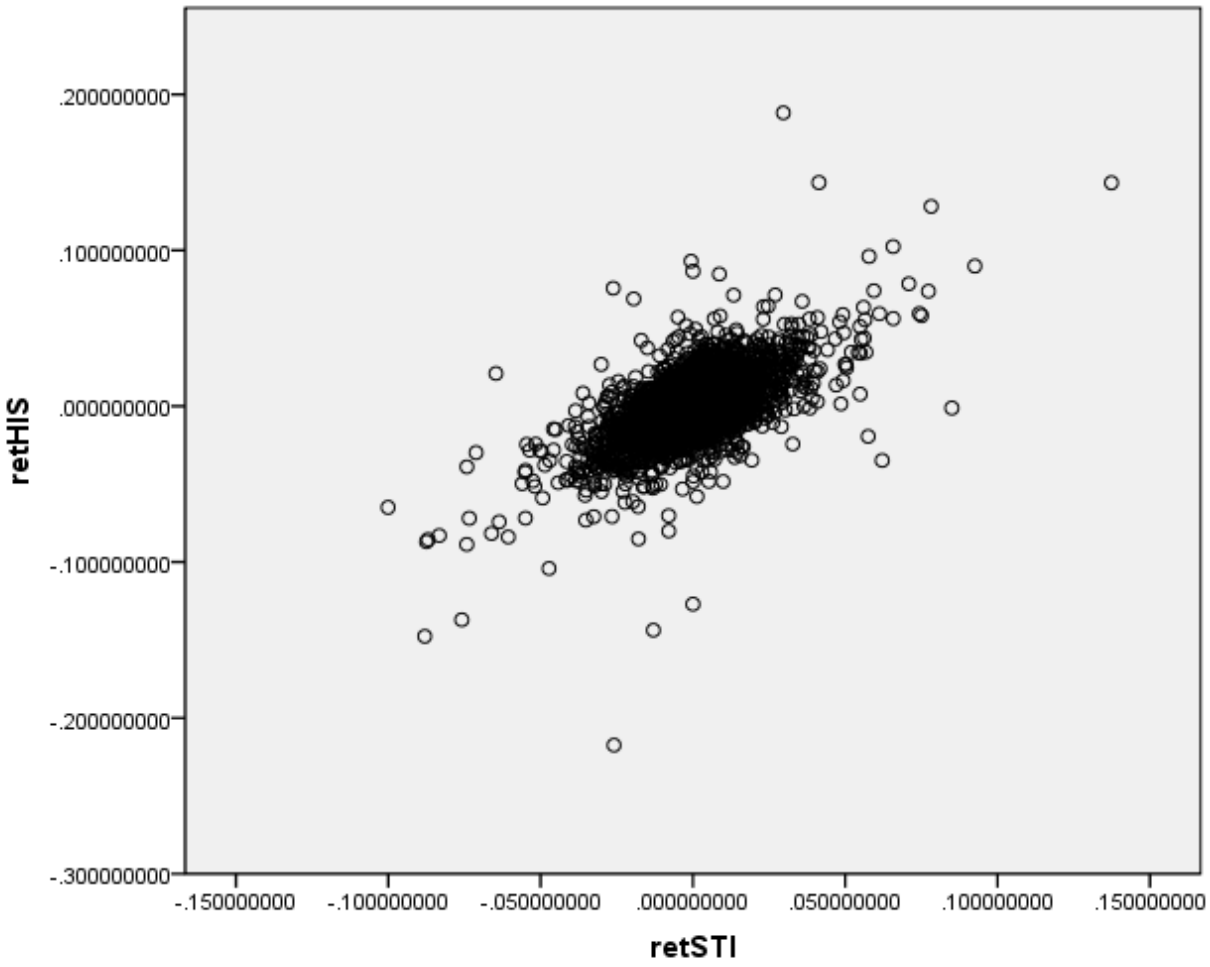
$$\overline{XY} = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i = \frac{1}{N}\sum_{i=1}^{N} X_i \{Y=1\} + \frac{1}{N}\sum_{i=1}^{N} X_i \{Y=0\} = \frac{1}{N}\sum_{i=1}^{N} X_i \{Y=1\} = \bar{X}_{Y=1}.$$

This is because the value of anything times zero is itself zero, so the term $\sum_{i=1}^{n} X_i \{Y=0\}$ drops out. While it is easy to see how this additional information is valuable when Y is a dummy variable, it is a bit more difficult to see that it is valuable when Y is a continuous variable – why might we want to look at the multiplied value, $X \cdot Y$?

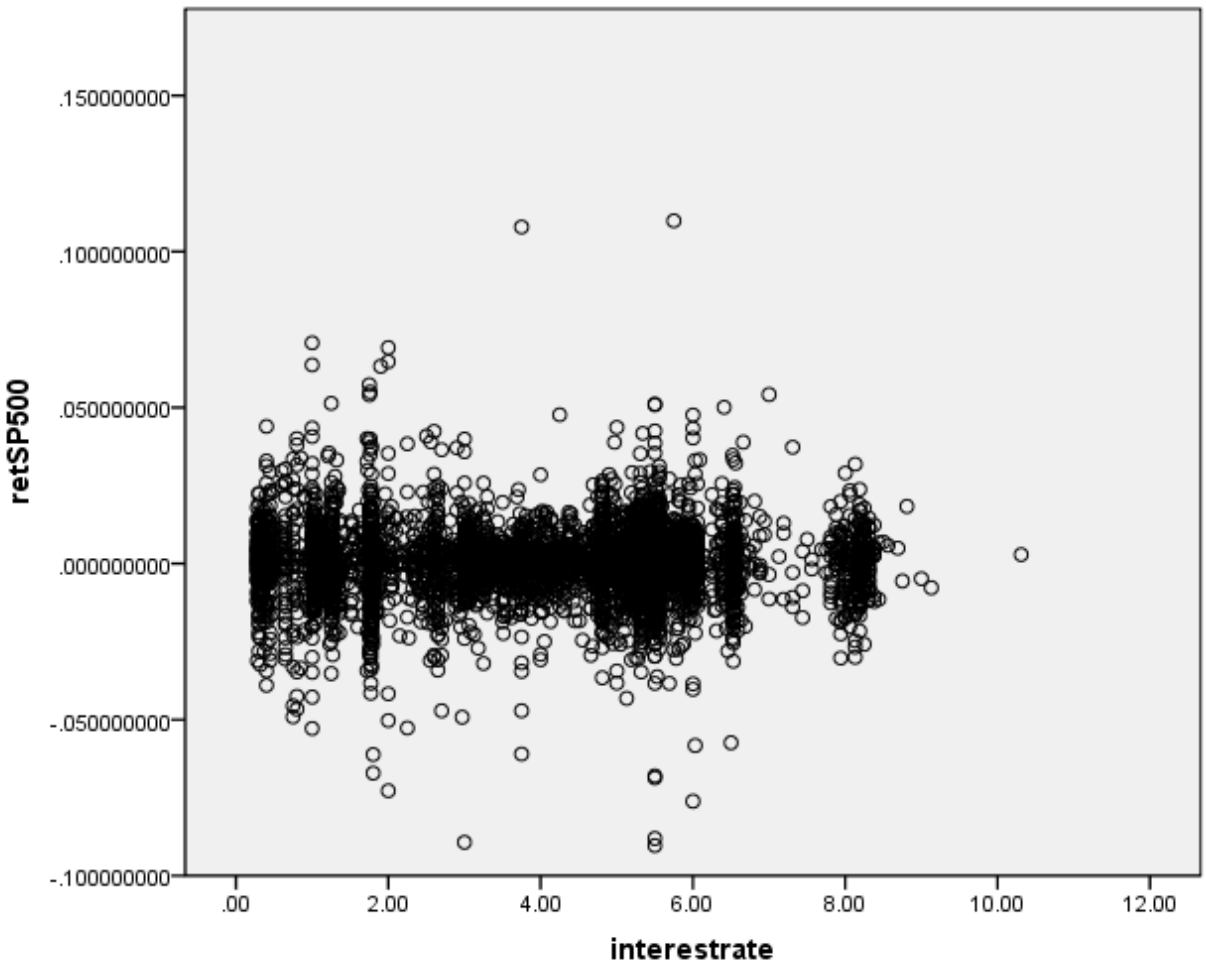
Use Your Eyes

We are accustomed to looking at graphs that show values of two variables and trying to discern patterns. Consider these two graphs of financial variables.

This plots the returns of Hong Kong's Hang Seng index against the returns of Singapore's Straits Times index (over the period from Dec 29, 1989 to Sept 1, 2010)



This next graph shows the S&P 500 returns and interest rates (1-month Eurodollar) during Jan 2, 1990 – Sept 1, 2010.



You don't have to be a highly-skilled econometrician to see the difference in the relationships. It would seem reasonable to state that the Hong Kong and Singapore stock returns are closely linked; while US stock returns are not closely related to US interest rates. (Remember, in most economic applications we want to use stock returns not the level of the price or index; typically returns are $ln(P_t) - ln(P_{t-1})$.)

We want to ask, how could we measure these relationships? Since these two graphs are rather extreme cases, how can we distinguish cases in the middle? And then there is one farther, even more important question: how can we try to guard against seeing relationships where, in fact, none actually exist? The second question is the big one, which most of this course (and much of the discipline of statistics) tries to answer. But start with the first question.

How can we measure the relationship?

Correlation measures how/if two variables move together.

Recall from above that we looked at the average of $X \cdot Y$ when Y was a dummy variable taking only the values of zero or one. Return to the case where Y is not a dummy but is a continuous variable just like X. It is still useful to find the average of $X \cdot Y$ even in the case where Y is from a continuous distribution and can take any value, $\overline{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$. It is a bit more useful if we re-write X and Y as differences from their means, so finding:

$$\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right).$$

This is the covariance, which is denoted cov(X,Y) or σ_{XY}.

A positive covariance shows that when X is above its mean, Y tends to also be above its mean (and vice versa) so either a positive number times a positive number gives a positive or a negative times a negative gives a positive.

A negative covariance shows that when X is above its mean, Y tends to be below its mean (and vice versa). So when one is positive the other is negative, which gives a negative value when multiplied.

A bit of math (extra):

$$\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) =$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(X_i Y_i - \bar{X}Y_i - X_i\bar{Y} + \bar{X}\bar{Y}\right) =$$

$$\frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \frac{1}{N}\sum_{i=1}^{N}\bar{X}Y_i - \frac{1}{N}\sum_{i=1}^{N} X_i\bar{Y} + \frac{1}{N}\sum_{i=1}^{N}\bar{X}\bar{Y} =$$

$$\frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}\frac{1}{N}\sum_{i=1}^{N} Y_i - \bar{Y}\frac{1}{N}\sum_{i=1}^{N} X_i + \bar{X}\bar{Y}\frac{1}{N}\sum_{i=1}^{N} 1$$

=

$$\frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y} - \bar{Y}\bar{X} + \bar{X}\bar{Y} =$$

$$\frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y}$$

(a strange case because it makes FOIL look like just FL!)

Covariance is sometimes scaled by the standard deviations of X and Y in order to eliminate problems of measurement units, so the correlation is:

$$\frac{\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sigma_X\sigma_Y}=\frac{\sigma_{XY}}{\sigma_X\sigma_y}=\rho_{XY}$$ or Corr(X,Y),

where the Greek letter "rho" denotes the correlation coefficient. With some algebra you can show that ρ is always between negative one and positive one; $-1\le\rho_{XY}\le 1$.

Two variables will have a perfect correlation if they are identical; they would be perfectly inversely correlated if one is just the negative of the other (assets and liabilities, for example). Variables with a correlation close to one (in absolute value) are very similar; variables with a low or zero correlation are nearly or completely unrelated.

Sample covariances and sample correlations

Just as with the average and standard deviation, we can estimate the covariance and correlation between any two variables. And just as with the sample average, the sample covariance and sample correlation will have distributions around their true value.

Go back to the case of the Hang Seng/Straits Times stock indexes. We can't just say that when one is big, the other is too. We want to be a bit more precise and say that when one is above its mean, the other tends to be above its mean, too. We might additionally state that, when the standardized value of one is high, the other standardized value is also high. (Recall that the standardized value of one variable, X, is $Z_{X,i}=\frac{X_i-\bar{X}}{s_X}$, and the standardized value of Y is $Z_{Y,i}=\frac{Y_i-\bar{Y}}{s_Y}$.)

Multiplying the two values together, $Z_{X,i}Z_{Y,i}$, gives a useful indicator since if both values are positive then the multiplication will be positive; if both are negative then the multiplication will again be positive. So if the values of Z_X and Z_Y are perfectly linked together then multiplying them together will get a positive number. On the other hand, if Z_X and Z_Y are oppositely related, so whenever one is positive the other is negative, then multiplying them together will get a negative number. And if Z_X and Z_Y are just random and not related to each other, then multiplying them will sometimes give a positive and sometimes a negative number.

Sum up these multiplied values and get the (population) correlation, $\frac{1}{N}\sum_{i=1}^{N}Z_{X,i}Z_{Y,i}$.

This can be written as $\frac{1}{N}\sum_{i=1}^{N}Z_{X,i}Z_{Y,i}=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i-\bar{X}}{s_x}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)=\frac{1}{N}\frac{1}{s_xs_Y}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)$. The population

correlation between X and Y is denoted ρ_{XY}; the sample correlation is r_{XY}. Again the difference is whether you divide by N or (N – 1). Both correlations are always between -1 and +1; $-1 \le \rho \le 1; -1 \le r \le 1$.

We often think of drawing lines to summarize relationships; the correlation tells us something about how well a line would fit the data. A correlation with an absolute value near 1 or -1 tells us that a line (with either positive or negative slope) would fit well; a correlation near zero tells us that there is "zero relationship."

The fact that a negative value can infer a relationship might seem surprising but consider for example poker. Suppose you have figured out that an opponent makes a particular gesture when her cards are no good – you can exploit that knowledge, even if it is a negative relationship. In finance, if a fund manager finds two assets that have a strong negative correlation, that one has high returns when the other has low returns, then again this information can exploited by taking offsetting positions.

You might commonly see a "covariance matrix" if you were working with many variables; the matrix shows the covariance (or correlation) between each pair. So if you have 4 variables, named (unimaginatively) X1, X2, X3, and X4, then the covariance matrix would be:

	X1	X2	X3	X4
X1	σ_{11}			
X2	σ_{21}	σ_{22}		
X3	σ_{31}	σ_{32}	σ_{33}	
X4	σ_{41}	σ_{42}	σ_{34}	σ_{44}

Where the matrix is "lower triangular" because cov(X,Y)=cov(Y,X) [return to the formulas if you're not convinced] so we know that the upper entries would be equal to their symmetric lower-triangular entry (so the upper triangle is left blank since the entries would be redundant). And we can also show [again, a bit of math to try on your own] that cov(X,X) = var(X) so the entries on the main diagonal are the variances.

If we have a lot of variables (15 or 20) then the covariance matrix can be an important way to easily show which ones are tightly related and which ones are not.

As a practical matter, sometimes perfect (or very high) correlations can be understood simply by definition: a survey asking "Do you live in a city?" and "Do you live in the countryside?" will get a very high negative correlation between those two questions. A firm's Assets and Liabilities ought to be highly correlated. But other correlations can be caused by the nature of the sampling.

Higher Moments

The third moment is usually measured by skewness, which is a common characteristic of financial returns: there are lots of small positive values balanced by fewer but larger negative values. Two portfolios could have the same average return and same standard deviation, but if one is not symmetric distribution (so has a non-zero skewness) then it would be important to understand this risk.

The fourth moment is kurtosis, which measures how fat the tails are, or how fast the probabilities of extreme values die off. Again a risk manager, for example, would be interested in understanding the differences between a distribution with low kurtosis (so lots of small changes) versus a distribution with high kurtosis (a few big changes).

If these measures are not perfectly clear to you, don't get frustrated – it is difficult, but it is also very rewarding. As the Financial Crisis has shown, many top risk managers at name-brand institutions did not understand the statistical distributions of the risks that they were taking on. They plunged the global economy into recession and chaos because of it.

These are called "moments" to reflect the origin of the average as being like weights on a lever or "moment arm". The average is the first moment, the variance is the second, skewness is third, kurtosis is fourth, etc. If you take a class using Calculus to go through Probability and Statistics, you will learn moment-generating functions.

More examples of correlation:

It is common in finance to want to know the correlation between returns on different assets.

First remember the difference between the returns and the level of an asset or index!

An investment in multiple assets, with the same return but that are uncorrelated, will have the same return but with less overall risk. We can show this on Excel; first we'll do random numbers to show the basic idea and then use specific stocks.

How can we create normally-distributed random numbers in Excel? RAND() gives random numbers between zero and one; NORMSINV(RAND()) gives normally distributed random numbers. (If you want variables with other distributions, use the inverse of those distribution functions.) Suppose that two variables each have returns given as 2% + a normally-distributed random number; this is shown in Excel sheet, lecturenotes3.xls

With finance data, we use the return not just the price. This is because we assume that investors care about returns per dollar not the level of the stock price.

Important Questions

- When we calculate a correlation, what number is "big"? Will see random errors – what amount of evidence can convince us that there is really a correlation?

- When we calculate conditional means, and find differences between groups, what difference is "big"? What amount of evidence would convince us of a difference?

Example:

Mazar, Amir, Ariely (2005) "Dishonesty of Honest People" [SSRN-id979648.pdf, available online]

Students solve math problems and report how many, of 20, were solved (offered a small reward for success). Here is a sample question: **Which 2 numbers add to 10**? You can see that finding the answer is tedious but doesn't require advanced mathematical knowledge.

1.69	1.82	2.91
4.67	4.81	3.05
5.82	5.06	4.28
6.36	5.19	4.57

In one setup, the students first threw out the answer sheet and then just said how many they'd solved; in the other setup they handed over the sheet to be checked – so it was easier to cheat in the first case. Students who had to hand in the sheet reported solving an average of 3.1 out of 20 problems in the short time given; students who threw out the sheet reported 4.2.

Are people more dishonest, when given a chance to be? Really? What information do we need, to be more confident about our knowledge? Ariely did another study looking at whether wearing counterfeit sunglasses made people more likely to cheat.

To answer these, we need to think about randomness – in other perceptual problems, what would be called noise or blur.

Learning Outcomes (from CFA exam Study Session 2, Quantitative Methods)

Students will be able to:

- calculate and interpret relative frequencies, given a frequency distribution, and describe the properties of a dataset presented as a histogram;

- define, calculate, and interpret measures of central tendency, including the population mean, sample mean, median, and mode;

- define, calculate, and interpret measures of variation, including the population standard deviation and the sample standard deviation;

- define and interpret the covariance and correlation;

- define a random variable, an outcome, an event, mutually exclusive events, and exhaustive events;

- distinguish between dependent and independent events;

Probability

Beyond presenting some basic measures such as averages and standard deviations, we want to try to understand how much these measures can tell us about the larger world. How likely is it, that we're being fooled, into thinking that there's a relationship when actually none exists? To think through these questions we must consider the logical implications of randomness and often use some basic statistical distributions (discrete or continuous).

Think Like a Statistician

The basic question that a Statistician must ask is "How likely is it, that I'm being fooled?" Once we accept that the world is random (rather than a manifestation of some god's will), we must decide how to make our decisions, knowing that we cannot guarantee that we will always be right. There is some risk that the world will seem to be one way, when actually it is not. The stars are strewn randomly across the sky but some bright ones seem to line up into patterns. So too any data might sometimes line up into patterns.

Statisticians tend to stand on their heads and ask, suppose there were actually no relationship? (Sometimes they ask, "suppose the conventional wisdom were true?") This statement, about "no relationship," is called the **Null Hypothesis**, sometimes abbreviated as H_0. The Null Hypothesis is tested against an **Alternative Hypothesis**, H_A.

Before we even begin looking at the data we can set down some rules for this test. We know that there is some probability that nature will fool me, that it will seem as though there is a relationship when actually there is none. The statistical test will create a model of a world where there is actually no relationship and then ask how likely it is that we could see what we actually see, "How likely is it, that I'm being fooled?" What if there were actually no relationship, is there some chance that I could see what I actually see?

Randomness in Sports

As an example, consider sports events. As any sports fan knows, a team or individual can get lucky or unlucky. The baseball World Series, for example, has seven games. It is designed to ensure that, by the end, one team or the other wins. But will the better team always win?

First make a note about subjectivity: if I am a fan of the team that won, then I will be convinced that the better team won; if I'm a fan of the losing team then I'll be certain that the better team got unlucky. But fans of each team might agree, if they discussed the question before the Series were played, that luck has a role.

Will the better team win? Clearly a seven-game Series means that one team or the other will win, even if they are exactly matched (if each had precisely a 50% chance of winning). If two representatives tossed a coin in the air seven times, then one or the other would win at least four tosses – maybe even more. We can use a computer to simulate seven

coin-tosses by having it pick a random number between zero and one and defining a "win" as when the random number is greater than 0.5.

Or instead of having a computer do it, we could use a bit of statistical theory.

Some math

Suppose we start with just one coin-toss or game (baseball uses 7 games to decide a champion; football uses just one). Choose to focus on one team so that we can talk about "win" and "loss". If this team has a probability of winning that is equal to *p*, then it has a probability of losing equal to *(1-p)*. So even if p, the probability of winning, is equal to 0.6, there is still a 40% chance that it could lose a single game. In fact unless the probability of winning is 100%, there is some chance, however remote, that the lesser team will win.

What about if they played two games? What are the outcomes? The probability of a team winning both games is $p*p = p^2$. If the probability were 0.5 then the probability of winning twice in a row would be 0.25.

A table can show this:

	Win Game 1 {p}	Lose Game 1 {1-p}
Win Game 2 {p}	outcome: W,W	L,W
Lose Game 2 {1-p}	W,L	L,L

This is a fundamental fact about how probabilities are represented mathematically: if the probabilities are not related (i.e. if the tossed coin has no memory) then the probability of both events happening is found my multiplying the probabilities of each individual outcome. (What if they're not unrelated, you may ask? What if the first team that wins gets a psychological boost in the next so they're more likely to win the second game? Then the math gets more complicated – we'll come back to that question!)

The math notation for two events, call them A and B, both happening is:

$$\Pr\{A \text{ and } B\} = \Pr\{A \cap B\}$$

The fundamental fact of independence is then represented as:

$$\Pr\{A \cap B\} = \Pr\{A\}\Pr\{B\} \quad \text{if } A \text{ and } B \text{ are independent}$$

where we use the term "independent" for when there is no relationship between them.

The probability that a team could lose both games is (1-p)*(1-p) = $(1-p)^2$. The probability that the teams could split the series (each wins just one) is p*(1-p) + (1-p)*p = 2p(1-

p). There are two ways that each team could win just one game: either the series splits (Win,Loss) or (Loss,Win).

For three games the outcomes become more complicated: now there are 8 combinations of win and loss:

(W,W,W)	(W,W,L)	(W,L,W)	(L,W,W)	(W,L,L)	(L,W,L)	
p*p*p	p*p*(1-p)	p*(1-p)p	(1-p)p*p	p(1-p)(1-p)	(1-p)p(1-p)	

and the probabilities are in the row below.

The team will win the series in any of the left-most 4 outcomes so its overall probability of winning the series is

$$p^3 + 3p^2(1-p)$$

while its probability of losing the series is

$$3p(1-p)^2 + (1-p)^3.$$

Clearly if p is 0.5 so that p=(1-p) then the chances of either team winning the three-game series are equal. If the probabilities are not equal then the chances are different, but as long as there is a probability not equal to one or zero (i.e. no certainty) then there is a chance that the worse team could win.

If you keep on working out the probabilities for longer and longer series you might notice that the coefficients and functional forms are right out of Pascal's Triangle. This is your first notice of just how "normal" the Normal Distribution is, in the sense that it jumps into all sorts of places where you might not expect it. The terms of Pascal's Triangle begin (as N becomes large) to form a normal distribution! We'll come back to this again...

Terms and Definitions

Some basics: a sample space is the entire list of possible outcomes (can be whole long list or even mathematical sets such as real numbers); events are subsets of the sample space. Simple event is a single outcome (one dice comes up 6); a compound event is several outcomes (both dice come up 6). Notate an event as A. The complement of the event is the set of all events that are not in A; this is A'.

The events must be **mutually exclusive and exhaustive**, so a good deal of the hard work in probability is just figuring out how to list all of the events.

Mutually exclusive means that the events must be clearly defined so that the data observed can be classified into just one event. Exhaustive means that every possible data

observed must fit into some event. The "mutually exclusive" part means that probabilities can be added up, so that if the probability of rolling a "1" on a dice is 1/6 and the probability of rolling a 6 is 1/6, then the probability of rolling either a 1 or 6 is 2/6 = 1/3. The "exhaustive" part of defining the events means that the sum of all the events must equal one.

For example, suppose we roll two dice. We might want to think of "die #1 comes up as 6" as one event [in English, the singular of "dice" is "die" – how morbid gambling can be!]. But the other die can have 6 different values without changing the value of the first die. So a better list of events would be the integers from 2 to 12, the sum of the dice values – with the note that there are many ways of achieving some of the events (a 7 is a 6 &1 or a 5&2, or 4&3, or 3&4, or 2&5, or 1&6) while other events have only one path (each die comes up 6 to make 12).

A **sample space** is the set of all possible events. The sum of the probability of all of the events in the sample space is equal to one. There is a 100% chance that something happens (provided we've defined the sample space correctly). So if a lottery brags that there is a 2% chance that "you might be a winner!" this is equivalent to stating that there is a 98% chance that you'll lose.

Events have **probability**; this must lie between zero and one (inclusive); so $0 \le P \le 1$. The probability of all of the events in the sample space must sum to one. This means that the probability of an event and its complement must sum to one: $P\{A\} + P\{A'\} = 1$.

Probabilities come from empirical results (relative frequency approach) or the classical (a priori or postulated) assignment or from subjective beliefs that people have.

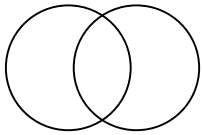
In empirical approach, the **Law of Large Numbers** is important: as the number of identical trials increases, the estimated frequency approaches its theoretical value. You can try flipping coins and seeing how many come up heads *(flip a bunch at a time to speed up the process)*; it should be 50%.

We are often interested in finding the probability of two events both happening; this is the "**intersection**" of two events; the logical "and" relationship; two things both occurring. In the PUMS data we might want to find how many females have a college degree; in poker we might care about the chance of an opponent having an ace as one of her hole cards and the dealer turning up a king. We notate the intersection of A and B as $A \cap B$ and want to find $P\{A \cap B\}$. In SPSS this is notated with "&".

The "**union**" of two events is the logical "or" so it is either of two events occurring; this is $A \cup B$ so we might consider $P\{A \cup B\}$ or, in SPSS, "|". In the PUMS data we might want to combine people who report themselves as having race "black" with those who report "black – white". In cards, it is the probability that any of my 3 opponents has a better hand.

Married people can buy life insurance policies that pay out either when the first person dies or after both die – logical *and* vs *or*.

Venn Diagrams (Ballantine)



General Law of Addition

$$P\{A\cup B\} = P\{A\} + P\{B\} - P\{A\cap B\}$$

and so $P\{A\cap B\} = P\{A\} + P\{B\} - P\{A\cup B\}$

Mutually Exclusive (Special Law of Addition),

If $A\cap B = \phi$ then $P\{A\cap B\} = 0$ and $P\{A\cup B\} = P\{A\} + P\{B\}$

Conditional Probability

$P\{A|B\} = \dfrac{P\{A\cap B\}}{P\{B\}}$ if $P\{B\} \neq 0$. See Venn Diagram.

Independent Events

A is independent of B if and only if $P\{A|B\} = P\{A\}$

If we have multiple random variables then we can consider their **Joint Distribution**: the probability associated with each outcome in both sample spaces. So a coin flip has a simple discrete distribution: a 50% chance of heads and a 50% chance of tails. Flipping 2 coins gives a joint distribution: a 25% chance of both coming up heads, a 25% chance of both coming up tails, and a 50% chance of getting one head and one tail.

The probability of multiple independent events is found by multiplying the probabilities of each event together. So the chance of rolling two 6 on two dice is $\frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$. The probability of getting to the computer lab on the 6th floor of NAC from the first floor, without having to walk up a broken escalator, can be found this way too. Suppose the probability of an escalator not working is p; then the probability of it working is $(1-p)$ and the probability of five escalators each working is $(1-p)^5$. So even if the probability of a breakdown is small (5%), still the probability of having every escalator work is just $(1-5\%)^5 = (95\%)^5 = (0.95)^5 = \left(\frac{95}{100}\right)^5 = 0.7738 = 77.38\%$ so this implies that you'd expect to walk more than once a week.

A simple representation of the joint distribution of two coin flips is a table:

	coin 1 Heads	coin 1 Tails
coin 2 Heads	H,H at 25%	H,T at 25%
coin 2 Tails	T,H at 25%	T,T at 25%

Where, since the outcomes are independent, we can just multiply the probabilities.

The Joint Distribution tells the probabilities of all of the different outcomes. A **Marginal Distribution** answers a slightly different question: given some value of one of the variables, what are the probabilities of the other variables?

When the variables are independent then the marginal distribution does not change from the joint distribution. Consider a simple example of X and Y discrete variables. X takes on values of 1 or 2 with probabilities of 0.6 and 0.4 respectively. Y takes on values of 1, 2, or 3 with probabilities of 0.5, 0.3, and 0.2 respectively. So we can give a table like this:

	X=1 (60%)	X=2 (40%)
Y=1 (50%)	(1,1) at probability 0.3	(2,1) at probability 0.2
Y=2 (30%)	(1,2) at probability 0.18	(2,2) at probability 0.12
Y=3 (20%)	(1,3) at probability 0.12	(2,3) at probability 0.08

On the assumption that X and Y are independent. The probabilities in each box are found by multiplying the probability of each independent event.

If instead we had the two variables, A and B, not being independent then we might have a table more like this:

	A=1	A=2
B=1	(1,1) at probability 0.25	(2,1) at probability 0.13
B=2	(1,2) at probability 0.23	(2,2) at probability 0.12

B=3	(1,3) at probability 0.17	(2,3) at probability 0.1

We will examine the differences.

If we add up the probabilities along either rows or columns then we get the **marginal probabilities** (which we write in the *margins*, appropriately enough). Then we'd get:

	X=1 (60%)	X=2 (40%)	
Y=1 (50%)	(1,1) at probability 0.3	(2,1) at probability 0.2	0.5
Y=2 (30%)	(1,2) at probability 0.18	(2,2) at probability 0.12	0.3
Y=3 (20%)	(1,3) at probability 0.12	(2,3) at probability 0.08	0.2
	0.6	0.4	

Which just re-states our assumption that the variables are independent – and shows that, where there is independence, the probability of either variable alone does not depend on the value that the other variable takes on. In other words, knowing X does not give me any information about the value that Y will take on, and vice versa.

If instead we do this for the A,B case we get:

	A=1	A=2	
B=1	(1,1) at probability 0.25	(2,1) at probability 0.13	0.38
B=2	(1,2) at probability 0.23	(2,2) at probability 0.12	0.35
B=3	(1,3) at probability 0.17	(2,3) at probability 0.1	0.27
	0.65	0.35	

Where we double check that we've done it right by seeing that the sum of either of the marginals is equal to one (65% + 35% = 100% and 38% + 35% + 27% = 100%).

So the marginal distributions sum the various ways that an outcome can happen. For example, we can get A=1 in any of 3 ways: either (1,1), (1,2) or (1,3). So we add the probabilities of each of these outcomes to find the total chance of getting A=1.

But if we want to understand how A and B are related, it might be more useful to consider this as a prediction problem: would knowing the value that A takes on help me guess the value of B? Would knowing the value that B takes on help me guess the value of A?

These are abstract questions but they have vitally important real-life analogs. In airport security, is the probability that someone is a terrorist independent of whether they are Muslim? Is the probability that someone is pulled out of line for a thorough search independent of whether they are Muslim? *(The TSA might have different beliefs than you or me!)* In medicine, is the probability that someone gets cancer independent of whether they eat lots of vegetables? In economics, is the probability that someone defaults on their mortgage independent of the mortgage originator (Fannie, Freddie, mortgage broker, bank)? Is the probability of the country pulling out of recession independent of whether the Fed raises rates? In poker, if my opponent just raised the bid, what is the probability that her cards are better than mine?

For these questions we want to find the conditional distribution: what is the probability of some outcome, given a particular value for some other random variable?

Just from the phrasing of the question, you should be able to see that if the two variables are independent then the conditional distribution should not change from the marginal distribution – as is the case of X and Y. Flipping a coin does not help me guess the outcome of a roll of the dice. (Cheering in front of a sports game on TV does not affect the outcome, for another example – although plenty of people act as though they don't believe that!)

How do we find the conditional distribution? Take the value of the joint distribution and divide it by the marginal distribution of the relevant variable.

For example, suppose we want to find the probability of B outcomes, conditional on A=1. Since we know that A=1, there is no longer a 65% probability of A -- it happened. So we divide each joint probability by 0.65 so that the sum will be equal to 1. So the probabilities are now:

	A=1	A=2	
B=1	(1,1) at probability 0.25/.65	(2,1) at probability 0.13	0.38
B=2	(1,2) at probability 0.23/.65	(2,2) at probability 0.12	0.35
B=3	(1,3) at probability 0.17/.65	(2,3) at probability 0.1	0.27
	0.65/.65	0.35	

so now we get the conditional distribution:

	A=1	A=2	
B=1	(1,1) @ 0.3846	(2,1) at probability 0.13	0.38
B=2	(1,2) @ 0.3538	(2,2) at probability 0.12	0.35
B=3	(1,3) @ 0.2615	(2,3) at probability 0.1	0.27
		0.35	

We could do the same to find the conditional distribution of B, given that A=2:

	A=1	A=2	
B=1	(1,1) at probability 0.25	(2,1) @ 0.13/.35 =.3714	0.38
B=2	(1,2) at probability 0.23	(2,2) @ 0.12/.35 = .3429	0.35
B=3	(1,3) at probability 0.17	(2,3) @ 0.1/.35 = .2857	0.27
	0.65		

These conditional probabilities are denoted as $\Pr\{B|A=2\}$ for example. We could find the expected value of B given that A equals 2, $E[B|A=2]$, just by multiplying the value of B by its probability of occurrence, so $E[B|A=2]=(1\cdot.3714)+(2\cdot.3429)+(3\cdot.2857)$.

We could find the conditional probabilities of A given B=1 or given B=2 or given B=3. In those cases we would sum across the rows rather than down the columns.

More pertinently, we can get crosstabs on two variables, for example the wage by education. First I break wages into groups: less than $10,000 per year; then up to $50,000; up to $100,000; and greater than that. The R-output (see working_on_PUMS_2.R for details) is:

	No HS	HS	SmColl	AS	Bach	Adv
less than 10,000	56734	26279	17648	5167	9859	6684
10,001 - 50,000	4806	13147	9440	5080	7983	4155
50,001 - 100,000	538	3303	3250	2421	6380	5703
100,001+	78	380	592	370	2746	3571

But these are raw numbers of people not fractions – so divide by the total number of observations (easy in Excel or can be done in R, depending on your preference); I also show the marginal:

	No HS	**HS**	**SmColl**	**AS**	**Bach**	**Adv**	**Marginals**
less than 10,000	0.2890	0.1339	0.0899	0.0263	0.0502	0.0340	0.6233
10,001 - 50,000	0.0245	0.0670	0.0481	0.0259	0.0407	0.0212	0.2272
50,001 - 100,000	0.0027	0.0168	0.0166	0.0123	0.0325	0.0291	0.1100
100,001+	0.0004	0.0019	0.0030	0.0019	0.0140	0.0182	0.0394
Marginals	0.3166	0.2196	0.1576	0.0664	0.1374	0.1025	

These numbers are rough to interpret; the conditionals might be easier. So can ask, what is the likelihood of making particular levels of wage income, conditional on level of education? This divides each proportion by its column sum, its marginal. Note each column sums to 1.

Conditional on Education	**No HS**	**HS**	**SmColl**	**AS**	**Bach**	**Adv**
less than 10,000	0.9128	0.6096	0.5706	0.3963	0.3656	0.3323
10,001 - 50,000	0.0773	0.3050	0.3052	0.3896	0.2960	0.2066
50,001 - 100,000	0.0087	0.0766	0.1051	0.1857	0.2366	0.2835
100,001+	0.0013	0.0088	0.0191	0.0284	0.1018	0.1775

This shows that, of the people without a high school diploma, 91% have wage of $10,000 or less, while just 33% of people with an Advanced Degree make that little money. On the opposite end, just about 1/10 of 1% of people without a high school diploma make over $100k, while nearly 18% of people with an Advanced Degree make more than $100k.

The other conditional is asking, of people with wages above $100,000, what fraction have each degree? That table is found by dividing each row by its sum:

Conditional on Wage	No HS	HS	SmColl	AS	Bach	Adv
less than 10,000	0.4636	0.2147	0.1442	0.0422	0.0806	0.0546
10,001 - 50,000	0.1077	0.2947	0.2116	0.1139	0.1789	0.0931
50,001 - 100,000	0.0249	0.1530	0.1505	0.1121	0.2954	0.2641
100,001+	0.0101	0.0491	0.0765	0.0478	0.3549	0.4615

So this shows that, of people making more than $100,000 in wages, 46% of them have an Advanced Degree, another 35% have a Bachelor's Degree, while just 18% have fewer educational qualifications.

Both of these conditioning sets help understand how education and wages are interrelated – there is not necessarily one better than the other. (Also, not all of these are working people – there are children, retirees, and others not in the workforce. You can re-do the numbers for subsets, maybe people 25-55 would be a better choice?)

Conditional probabilities can also be calculated with what is called **Bayes' Theorem**:

$$P\{B|A\} = \frac{P\{A|B\} \cdot P\{B\}}{P\{A\}}.$$

This can be understood by recalling the definition of conditional probability, $P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}$, so $P\{B|A\} = \frac{P\{A \cap B\}}{P\{A\}}$, that the conditional probability equals the joint probability divided by the marginal probability.

The power of Bayes' Theorem can be understood by thinking about medical testing. Suppose a genetic test screens for some disease with 99% accuracy. Your test comes back positive – how worried should you be? The surprising answer is not 99% worried; in fact often you might be more than likely to be healthy! Suppose that the disease is rare so only 1 person in 1000 has it (so 0.1%). So out of 1000 people, one person has the disease and the test is 99% likely to identify that person. Out of the remaining 999 people, 1% will be misidentified as having the disease, so this is 9.99 – call it 10 people. So eleven people will test positive but only one will actually have the disease so the probability of having the disease given that the test comes up positive, $P\{sick|test+\}$, is $\frac{P\{test+|sick\}P\{sick\}}{P\{test+\}} = \frac{0.99 \cdot 0.001}{0.01} = .099$.

The test is not at all useless – it has brought down an individual's likelihood of being sick by orders of magnitude, from one-tenth of one percent to ten percent. But it's still not nearly as accurate as the "99%" label might imply.

Many healthcare providers don't quite get this and explain it merely as "don't be too worried until we do further tests." But this is one reason why broad-based tests can be very expensive and not very helpful. These tests are much more useful if we first narrow down the population of people who might have the disease. For example home pregnancy tests might be 99% accurate but if you randomly selected 1000 people to take the test, you'd find many false positives. Some of those might be guys (!) or women who, for a variety of reasons, are not likely to be pregnant. The test is only useful as one element of a screen that gets progressively finer and finer.

Counting Rules

If A can occur as N_1 events and B can be N_2 events then the sample space is $N_1 \cdot N_2$ (visualize a contingency table with N_1 rows and N_2 columns).

Factorials: If there are N items then they can be arranged in $N! = (n)(n-1)(n-2)\ldots(1) = \prod_{i=0}^{N-1}(N-i)$ ways.

Permutations: **n** events that can occur in **r** items (where order is important) have a total of $nPr = \frac{n!}{(n-r)!}$ possible outcomes.

Combinations: n events that can occur in r items (where order is not important) have $nCr = \frac{n!}{r!(n-r)!}$ possible outcomes – just the permutation divided by r! to take care of the multiple ways of ordering.

So to apply these, consider computer passwords (see NYTimes article below).

The article reports:

> Mr. Herley, working with Dinei Florêncio, also at Microsoft Research, looked at the password policies of 75 Web sites. ... They reported that the sites that allowed relatively weak passwords were busy commercial destinations, including PayPal, Amazon.com and Fidelity Investments. The sites that insisted on very complex passwords were mostly government and university sites. What accounts for the difference? They suggest that "when the voices that advocate for usability are absent or weak, security measures become needlessly restrictive."

Consider the simple mathematics of why a government or university might want complex passwords. How many permutations are possible if passwords are 6 numerical digits? How many if passwords are 6 alphabetic or numeric characters? If the characters are alphabetic, numeric, and fifteen punctuation characters (, . _ - ? ! @ # $

% ^ & * ' ")? What if passwords are 8 characters? If each login attempt takes 1/100 of a second, how many seconds of "brute-force attack" does it take to access the account on average? If there is a penalty of 10 minutes after 3 unsuccessful login attempts, how long would it take to break in? (Of course, the article notes, if password requirements are so arcane that employees put their passwords on a Post-It attached to the monitor, then the calculations above are irrelevant.)

(for fun, here's another example of Joint/Marginal Distributions)

Tiger Mother Amy Chua in WSJ, Jan 8, 2011

> *A lot of people wonder how Chinese parents raise such stereotypically successful kids. They wonder what these parents do to produce so many math whizzes and music prodigies, what it's like inside the family, and whether they could do it too. Well, I can tell them, because I've done it. Here are some things my daughters, Sophia and Louisa, were never allowed to do:*
>
> • *attend a sleepover*
>
> • *have a playdate*
>
> • *be in a school play*
>
> • *complain about not being in a school play*
>
> • *watch TV or play computer games*
>
> • *choose their own extracurricular activities*
>
> • *get any grade less than an A*
>
> • *not be the No. 1 student in every subject except gym and drama*
>
> • *play any instrument other than the piano or violin*
>
> • *not play the piano or violin.*
>
> *I'm using the term "Chinese mother" loosely. I know some Korean, Indian, Jamaican, Irish and Ghanaian parents who qualify too. Conversely, I know some mothers of Chinese heritage, almost always born in the West, who are not Chinese mothers, by choice or otherwise. I'm also using the term "Western parents" loosely. Western parents come in all varieties.*

So you could go to PUMS and look at first-generation immigrants with parents from China, compare with other first-generation kids, see where are the Tiger Moms…

Lecture 2: Discrete and Continuous Random Variables

For any discrete random variable, the mean or expected value is:

$$E(X) = \mu = \sum_{i=1}^{N} x_i P(x_i)$$

and the variance is

$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 P(x_i)$ so the standard deviation is the square root.

These can be described by PDF or CDF – probability density function or cumulative distribution function. The PDF shows the probability of events; the CDF shows the cumulative probability of an event that is smaller than or equal to that event. The PDF is the derivative of the CDF.

Linear Transformations:

- If $Y = aX + b$ then Y will have mean $\mu_Y = a\mu_X + b$ and standard deviation $\sigma_Y = a\sigma_X$.
- If $Z = X + Y$ then $\mu_Z = \mu_X + \mu_Y$; $\sigma_Z = \sqrt{\sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}}$ (and if X and Y are independent then the covariance term drops out)

WARNING: These statements DO NOT work for non-linear calculations! The propositions above do NOT tell about when X and Y are multiplied and divided: the distributions of $X \cdot Y$ or X/Y are not easily found. Nor is $\ln X$, nor e^X. We might wish for a magic wand to make these work out simply but they **don't** in general.

Common Discrete Distributions:

Uniform

- depend on only upper and lower bound, so all events are in $[a, b]$
- mean is $\frac{a+b}{2}$; standard deviation is $\sqrt{\frac{[b-a+1]^2 - 1}{12}}$
- Many null hypotheses are naturally formulated as stating that some distribution is uniform: e.g. stock picks, names and grades, birth month and sports success, etc.

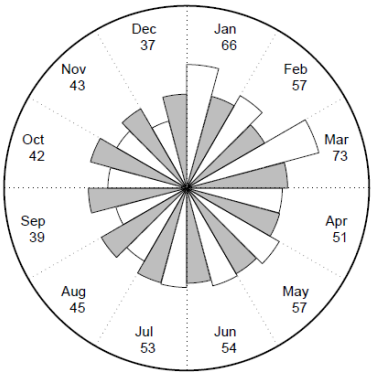


Figure 1: **Circular plot of the observed and expected number of AFL players' births.** The observed values are shown in white segments and the expected value in grey. The numbers around the outside of the plot are the observed number of births in each month. The expected number of births are based on national data.

from: Barnett, Adrian G. (2010) The relative age effect in Australian Football League players. Working Paper.

Although note that distribution of births is not quite uniform; certainly among animal species humans are unusual in that births are not overwhelmingly seasonal.

Benford's Law: not really a law but an empirical result about measurements, that looking at the first digit, the value 1 is much more common than 9 – the first digit is not uniformly distributed. Originally stated for tables of logarithms. Second digit is closer to uniform; third digit closer still, etc. See online Excel sheet. This is a warning that sometimes our intuition about how we might think numbers are distributed is actually wrong.

Question: Does the "shuffle" function on your music player distribute songs uniformly?

Bernoulli

- depend only on p, the probability of the event occurring
- mean is p; standard deviation is $\sqrt{p(1-p)}$
    - *Where is the maximum standard deviation? Intuition: what probability will give the most variation in yes/no answers? Or use calculus; note that has same maximum as p(1 – p) so take derivative of that, set to zero. Then hit your forehead with the palm of your hand, realizing that calculus gave you the same answer as simple intution.*
- Used for coin flips, dice rolls, events with "yes/no" answers: Was person re-employed after layoff? Did patient improve after taking the drug? Did company pay out to investors from IPO?

Binomial

- have n Bernoulli trials; record how many were 1 not zero
- $\mu = np$; $\sigma = \sqrt{np(1-p)}$

- These formulas are easy to derive from rules of linear combinations. If B_i are independent random variables with Bernoulli distributions, then what is the mean of $B_1 + B_2$? What is its std dev?
- What if this is expressed as a fraction of trials? Derive.

- what fraction of coin flips came up heads? What fraction of people were re-employed after layoff? What fraction of patients improved? What fraction of companies offereed IPOs?
- questions about opinion polls – the famous "plus or minus 2 percentage points"
  - get margin of error depending on sample size (*n*)
  - from above, figure that mean of the fraction of people who agree or support some candidate is p, the true value, with standard error of $\sqrt{p(1-p)}$.

Some students are a bit puzzled by two different sets of formulas for the binomial distribution – the standard deviation is listed as $\sqrt{np(1-p)}$ and $\sqrt{\frac{p(1-p)}{n}}$. Which is it?!

It depends on the units. If we measure the **number** of successes in **n** trials, then we multiply by **n**. If we measure the **fraction** of successes in **n** trials, then we don't multiply but divide.

Consider a simple example: the probability of a hit is 50% so $\sqrt{p(1-p)} = \sqrt{\frac{1}{2}\cdot\frac{1}{2}} = \sqrt{\frac{1}{4}} = \frac{1}{2}$. If we have 10 trials and ask, how many are likely to hit, then this should be a different number than if we had 500 trials. The standard error of the raw number of how many, of 10, hits we would expect to see, is $\sqrt{10}\cdot\frac{1}{2}$ which is 1.58, so with a 95% probability we would expect to see 5 hits, plus or minus 1.96*1.58 = 3.1 so a range between 2 and 8. If we had 500 trials then the raw number we'd expect to see is 250 with a standard error or $\sqrt{500}\cdot\frac{1}{2}$ = 11.18 so the 95% confidence interval is 250 plus or minus 22 so the range between 228 and 272. This is a bigger range (in absolute value) but a smaller part of the fraction of hits.

With 10 draws, we just figured out that the range of hits is (in fractions) from 0.2 to 0.8. With 500 draws, the range is from 0.456 to 0.544 – much narrower. We can get these latter answers if we take the earlier result of standard deviations and divide by n. The difference in the formula is just this result, since $\frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}$. You could think of this

as being analogous to the other "standard error of the average" formulas we learned, where you take the standard deviation of the original sample and divide by the square root of **n**.

Poisson

- model arrivals per time, assuming independent
- depends only on λ which is also mean
- PDF is $\frac{\lambda^x e^{-\lambda}}{x!}$
- model how long each line at grocery store is, how cars enter traffic, how many insurance claims

From Discrete to Continuous: an example of a very simple model (too simple)

Use computer to create models of stock price movements. What model? How complicated is "enough"?

Start really simple: Suppose the price were 100 today, and then each day thereafter it rises/falls by 10 basis points. What is the distribution of possible stock prices, after a year (250 trading days)?

Use Excel (not even R for now!)

First, set the initial price at 100; enter 100 into cell B2 (leaves room for labels). Put the trading day number into column A, from 1 to 250 (shortcut). In B1 put the label, "S".

Then label column C as "up" and in C2 type the following formula,

```
=IF(RAND()>0.5,1,0)
```

The "RAND()" part just picks a random number between 0 and 1 (uniformly distributed). If this is bigger than one-half then we call it "up"; if it's smaller then we call it "down". So that is the "=IF(statement, value-if-true, value-if-false)" portion. So it will return a 1 if the random number is bigger than one-half and zero if not.

Then label column D as "down" and in D2 just type

```
=1-C2
```

Which simply makes it zero if "up" is 1 and 1 if "up" is 0.

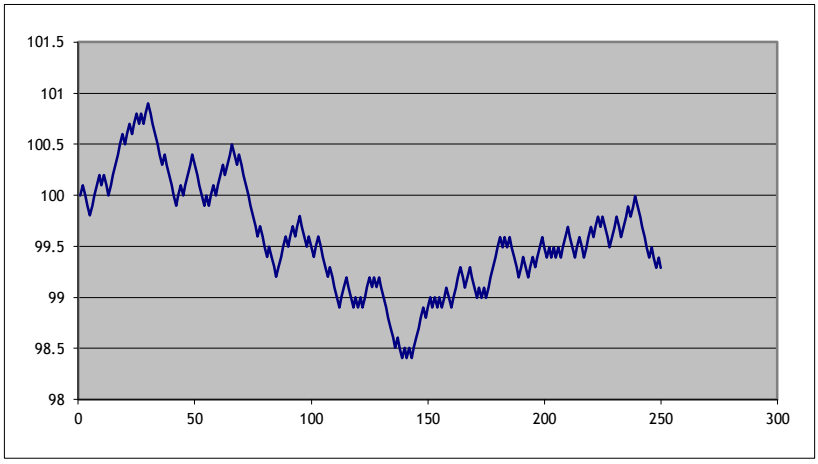
Then, in B3, put in the following formula,

```
=B2*(1+0.001*(C2-D2))
```

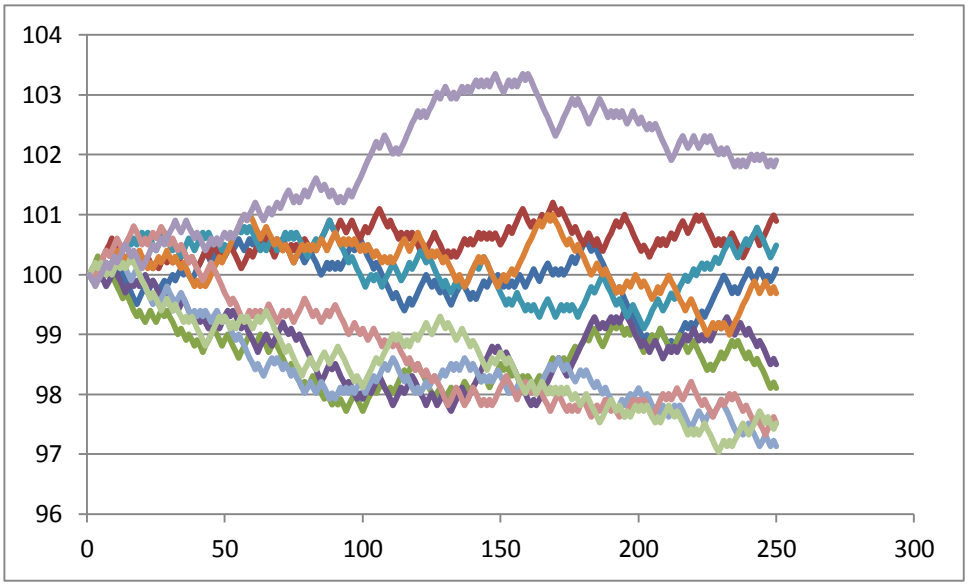
Copy and paste these into the remaining cells down to 250.

Of course this isn't very realistic but it's a start.

Then plot the result (highlight columns A&B, then "Insert\Chart\XY (Scatter)"); here's one of mine:



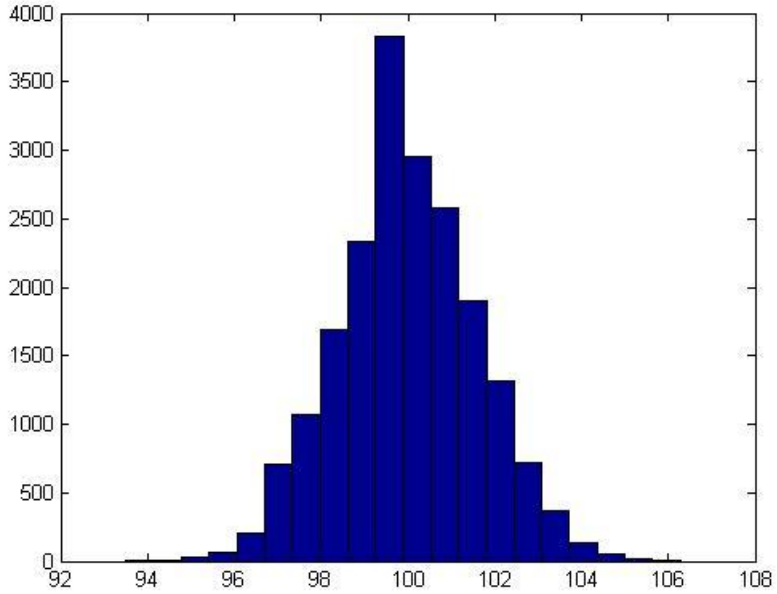
Here are 10 series (copied and pasted the whole S, "up," and "down" 10 times), see Excel sheet "*Lecturenotes2*".



We're not done yet; we can make it better. But the real point for now is to see the basic principle of the thing: we can simulate stock price paths as random trips.

The changes each day are still too regular – each day is 10 bps up or down; never constant, never bigger or smaller. That's not a great model for the middle parts. But the regularity within each individual series does not necessarily mean that the final prices (at step 250) are all that unrealistic.

I ran 2000 simulations; this is a histogram of the final price of the stock:



(If you're confident with you R knowledge, try writing that code!)

It shouldn't be a surprise that it looks rather normal (it is the result of a series of Bernoulli trials – that's what the Law of Large Numbers says should happen!).

With computing power being so cheap (those 2000 simulations of 250 steps took a few seconds) these sorts of models are very popular (in their more sophisticated versions).

It might seem more "realistic" if we thought of each of the 250 tics as being a portion of a day. ("Realistic" is a relative term; there's a joke that economists, like artists, tend to fall in love with their models.)

There are times (in finance for some option pricing models) when even this very simple model can be useful, because the fixed-size jump allows us to keep track of all of the possible evolutions of the price.

But clearly it's important to understand Bernoulli trials summing to Binomial distributions converging to normal distributions.

Continuous Random Variables

The PDF and CDF

Where discrete random variables would sum up probabilities for the individual outcomes, continuous random variables necessitate some more complicated math. When X is a continuous random variable, the probability of it being equal to any particular value is zero If X is continuous, there is a zero chance that it will be, say, 5 – it could be 4.99998 or 5.000001 and so on. But we can still take the area under the PDF by taking the limit of the sum, as the horizontal increments get smaller and smaller – the Riemann method, that you remember from Calculus. So to find the probability of X being equal to a set of values we integrate the PDF between those values, so

$$P\{a \le X \le b\} = \int_a^b p(x)\,dx\,.$$

The CDF, the probability of observing a value less than some parameter, is therefore the integral with $-\infty$ as the lower limit of integration, so $P\{X \le b\} = \int_{-\infty}^{b} p(x)\,dx\,.$

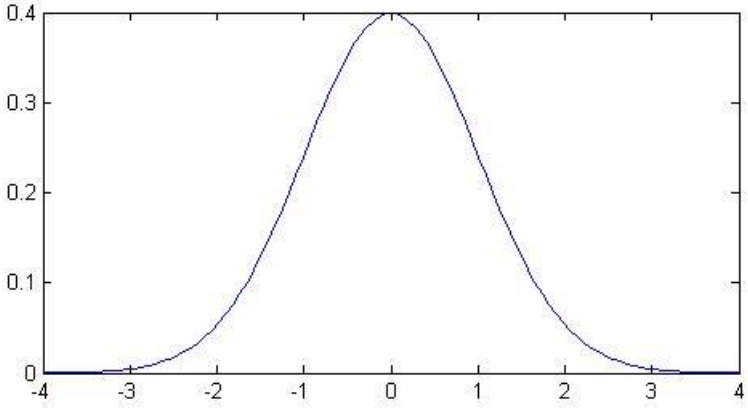
For this class you aren't required to use calculus but it's helpful to see why somebody might want to use it. *(Note that many of the statistical distributions we'll talk about come up in solving partial differential equations such as are commonly used in finance – so if you're thinking of a career in that direction, you'll want even more math!)*

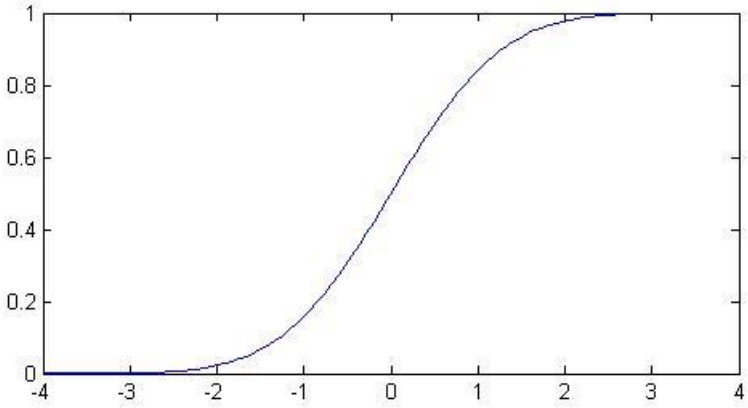
Normal Distribution

We will most often use the Normal Distribution – but usually the first question from students is "Why is that crazy thing normal?!!" You're not the only one to ask. Be patient, you'll see why; for now just remember e^{-x^2}.

In statistics it is often convenient to use a normal distribution, the bell-shaped distribution that arises in many circumstances. It is useful because the (properly scaled) mean of independent random draws of many other statistical distributions will tend toward a normal distribution – this is the Central Limit Theorem.

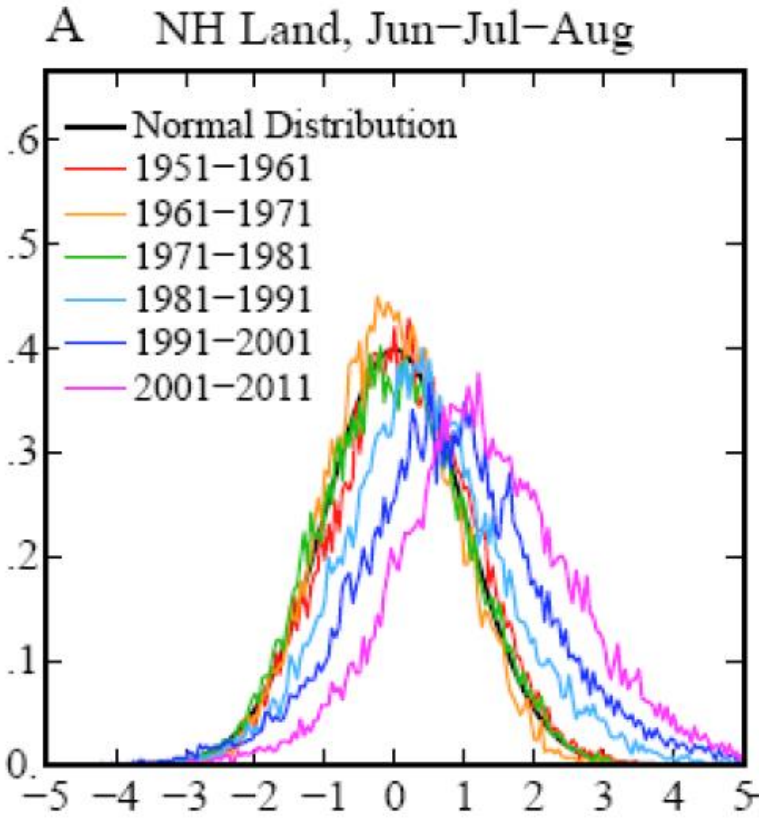
Some basic facts and notation: a normal distribution with mean μ and standard deviation σ is denoted $N(\mu,\sigma)$. (The variance is the square of the standard deviation, σ^2.) The Standard Normal distribution is when $\mu=0$ and $\sigma=1$; its probability density function (pdf) is denoted $pdf_N(x)$; the cumulative density function (CDF) is $cdf_N(x)$ or sometimes $Nor(x)$. This is a graph of the PDF (the height at any point) and CDF of the normal:





Example of using normal distributions:

A paper by Hansen, Sato, & Ruedy (2012) showed these decadal distributions of temperature anomalies:



This shows the rightward spread of temperature deviations. The x-axis is in standard deviations, which makes the various geographies easily comparable (a hot day in Alaska is different from a hot day in Oklahoma). The authors define extreme heat as more than 3 standard deviations above the mean and note that the probability of extreme heat days has risen from less than 1% to above 10%.

One of the basic properties of the normal distribution is that, if X is distributed normally with mean μ and standard deviation σ, then $Y = A + bX$ is also distributed normally, with mean $(A + b\mu)$ and standard deviation $b\sigma$. We will use this particularly when we "standardize" a sample: by subtracting its mean and dividing by its standard deviation, the result should be distributed with mean zero and standard deviation 1.

Oppositely, if we are creating random variables with a standard deviation, we can take random numbers with a N(0,1) distribution, multiply by the desired standard deviation, and add the desired mean, to get normal random numbers with any mean or standard deviation. In Excel, you can create normally distributed random numbers by using the RAND() function to generate uniform random numbers on [0,1], then NORMSINV(RAND()) will produce standard-normal-distributed random draws.

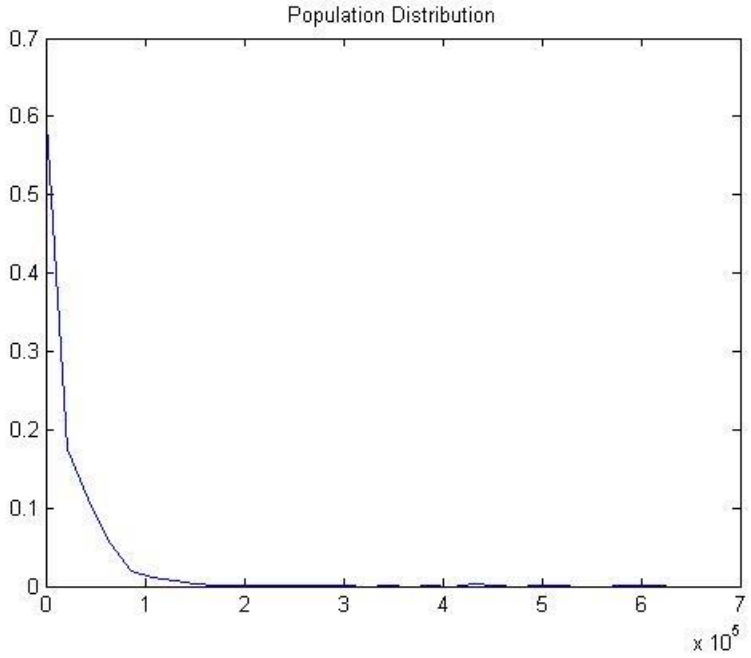
Motivation: Sample Averages are Normally Distributed

Before we do a long section on how to find areas under the normal distribution, I want to address the big question: Why we the heck would anybody ever want to know those?!?!

Consider a case where we have a population of people and we sample just a few to calculate an average. Before elections we hear about these types of procedures all of the time: a poll that samples just 1000 people is used to give information about how a population of millions of people will vote. These polls are usually given with a margin of error ("54% of people liked Candidate A over B, with a margin of error of plus or minus 2 percentage points"). If you don't know statistics then polls probably seem like magic. If you do know statistics then polls are based on a few simple formulas.

I have a dataset of about 206,639 people who reported their wage and salary to a particular government survey, the "Current Population Survey," the CPS. The true average of their wage and salaries was $19,362.62. (Not quite; the top income value is cut at $625,000 – people who made more are still just coded with that amount. But don't worry about that for now.) The standard deviation of the full 206,639 people is 39,971.91.

A histogram of the data shows that most people report zero (zero is the median value), which is reasonable since many of them are children or retired people. However some report incomes up to $625,000!



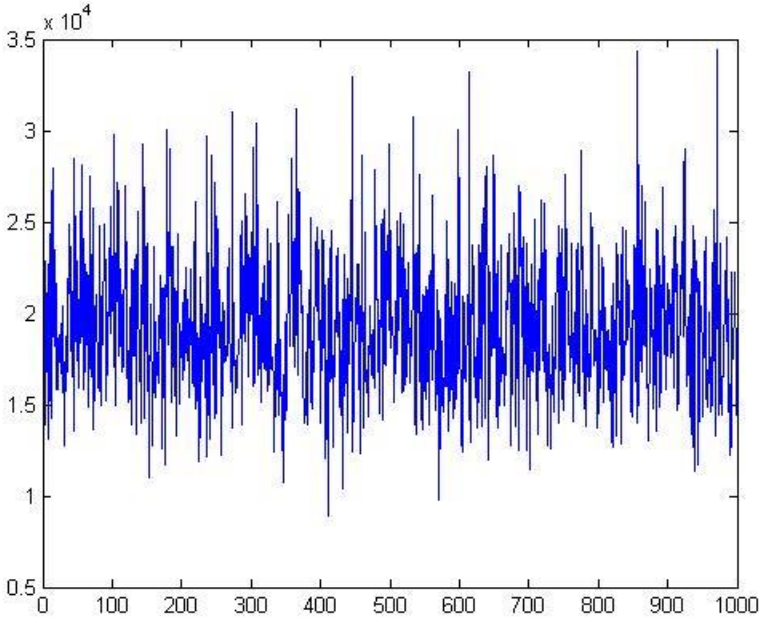
Taking an average of a population with such extreme values would seem to be difficult.

Suppose that I didn't want to calculate an average for all 206,639 people – I'm lazy or I've got a real old and slow computer or whatever. I want to randomly choose just 100 people and calculate the sample average. Would that be "good enough"?

Of course the first question is "good enough for what?" – what are we planning to do with the information?

But we can still ask whether the answer will be very close to the true value. In this case we know the true value; in most cases we won't. But this allows us to take a look at how the sampling works.

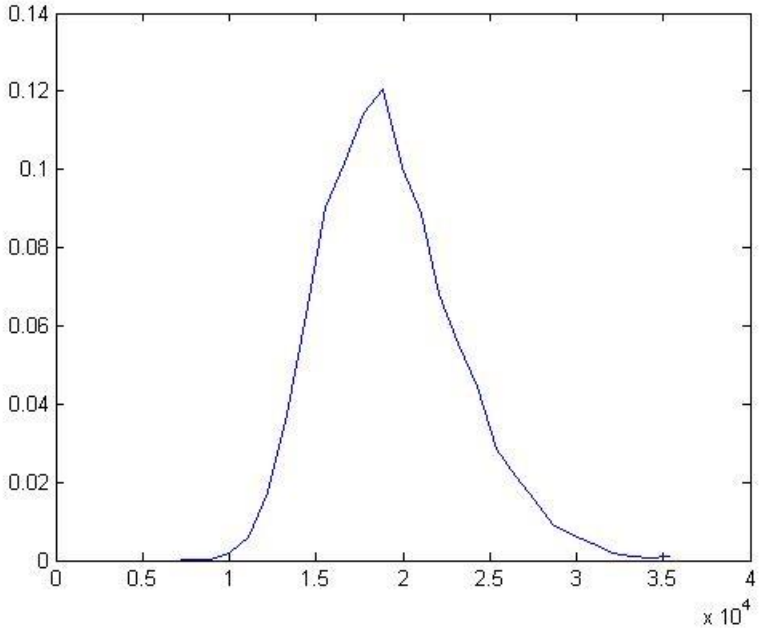
Here is a plot of values for 1000 different polls (each poll with just 100 people).



We can see that, although there are a few polls with averages as low almost 10,000 and a few with averages as high as 30,000, most of the polls are close to the true mean of $19,363.

In general the average of even a small sample is a good estimate of the true average value of the population. While a sample might pick up some extreme values from one side, it is also likely to pick extreme values from the other side, which will tend to balance out.

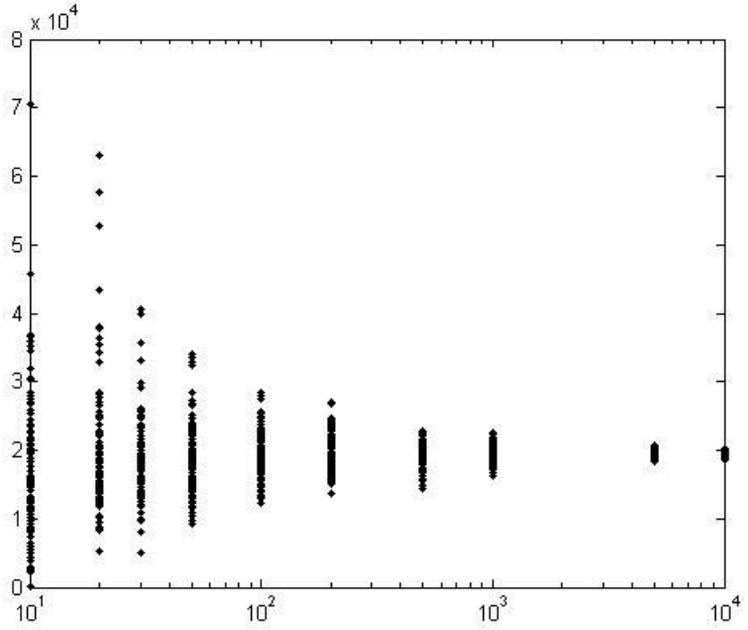
A histogram of the 1000 poll means is here:



This shows that the distribution of the sample means looks like a Normal distribution – another case of how "normal" and ordinary the Normal distribution is.

Of course the size of each sample, the number of people in each poll, is also important. Sampling more people gets us better estimates of the true mean.

This graph shows the results from 100 polls, each with different sample sizes.

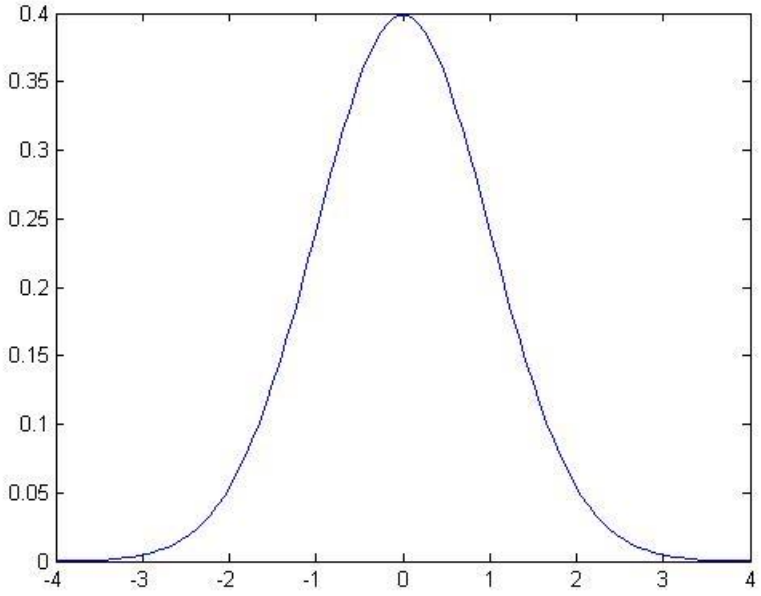


In the first set of 100 polls, on the left, each poll has just 10 people in it, so the results are quite varied. The next set has 20 people in each poll, so the results are closer to the true mean. By the time we get to 100 people in each poll (10^2 on the log-scale x-axis), the variation in the polls is much smaller.

Each distribution has a bell shape, but we have to figure out if there is a single invariant distribution or only a family of related bell-shaped curves.

If we subtract the mean, then we can center the distribution around zero, with positive and negative values indicating distance from the center. But that still leaves us with different scalings: as the graph above shows, the typical distance from the center gets smaller. So we divide by its standard deviation and we get a "Standard Normal" distribution.

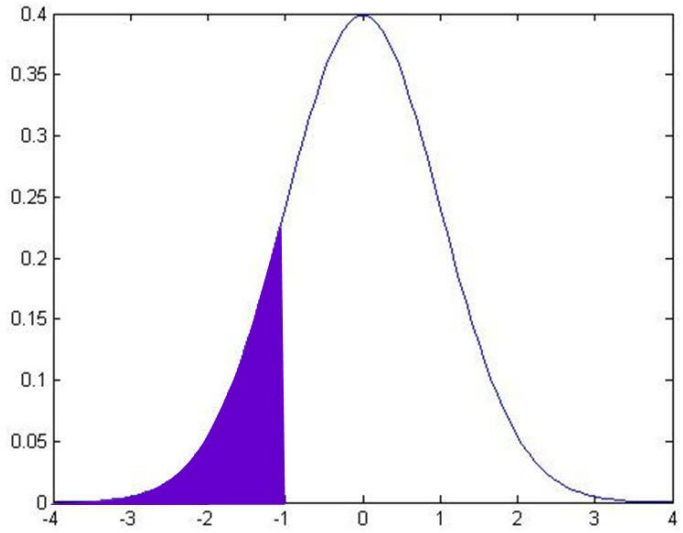
The Standard Normal graph is:



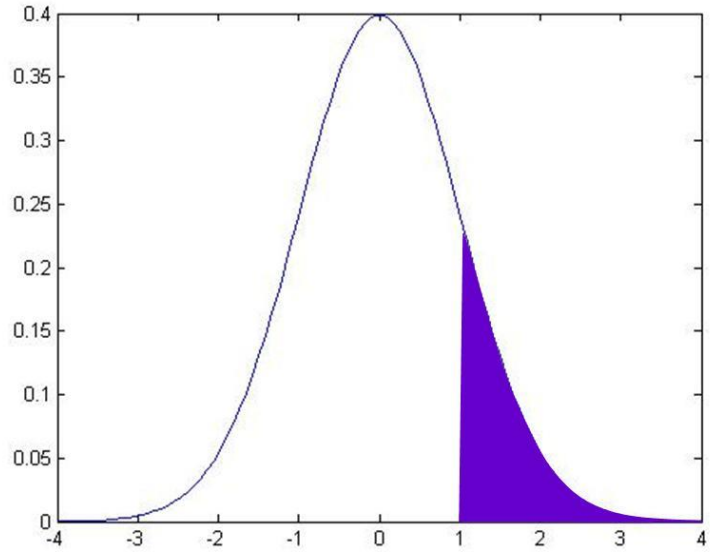
Note that it is symmetric around zero. Like any histogram, the area beneath the curve is a measure of the probability. The total area under the curve is exactly 1 (probabilities must add up to 100%). We can use the known function to calculate that the area under the curve, from -1 to 1, is 68.2689%. This means that just over 68% of the time, I will draw a value from within 1 standard deviation of the center. The area of the curve from -2 to 2 is 95.44997%, so we'll be within 2 standard deviations over 95.45% of the time.

It is important to be able to calculate areas under the Standard Normal. For this reason people used to use big tables (statistics textbooks still have them); now we use computers. But even the computers don't always quite give us the answer that we want, we have to be a bit savvy.

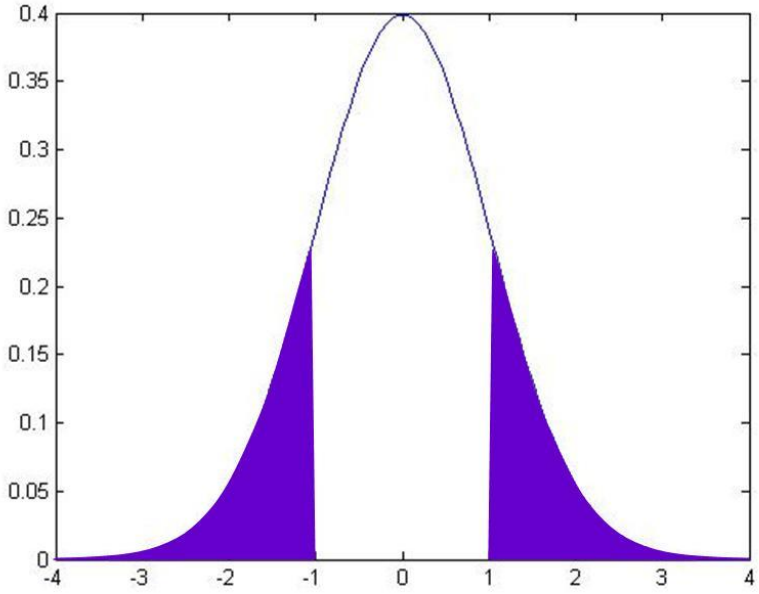
So the normal CDF of, say, -1, is the area under the pdf of the points to the left of -1:



This area is 15.87%. How can I use this information to get the value that I earlier told you, that the area in between -1 and 1 is 68.2689%? Well, we know two other things (more precisely, I know them and I wrote them just 3 paragraphs up, so you ought to know them). We know that the total area under the pdf is 100%. And we know that the pdf is symmetric around zero. This symmetry means that the area under the other tail, the area from +1 all the way to the right, is also 15.87%.



So to find the area in between -1 and +1, I take 100% and subtract off the two tail areas:



And this middle area is 100 – 15.87 – 15.87 = 68.26.

Sidebar: you can think of all of this as "adding up" without calculus. On the other hand, calculus makes this procedure much easier and we can precisely define the cdf as the integral,

$$cdf\left(Z\right)=\int_{-\infty}^{Z} pdf\left(x\right)dx$$

from negative infinity to some point Z, under the pdf: .

So with just this simple knowledge, you can calculate all sorts of areas using just the information in the CDF.

Hints on using Excel or R to calculate the Standard Normal cdf

Excel

Excel has both `normdist` and `normsdist`. For `normdist`, you need to tell it the mean and standard deviation, so use the function `normdist(X,mean,stdev,cumulative)`. For `normsdist` it assumes the mean is zero and standard deviation is one so you just use `normsdist(X)`. Read the help files to learn more. The final argument of the `normdist` function, "`Cumulative`" is a true/false: if true then it calculates the cdf (area to the left of X); if false it calculates the pdf. *[Personally, that's an ugly and non-intuitive bit of coding, but then again, Microsoft has no sense of beauty.]*

To figure out the other way – what X value gives me some particular probability, we use `norminv` or `normsinv`.

All of these commands are under "`Insert`" then "`Function`" then, under "`Select a Category`" choose "`Statistical`".

Google

Mistress Google knows all. When I google "Normal cdf calculator" I get a link to http://www.uvm.edu/~dhowell/StatPages/More_Stuff/normalcdf.html. This is a simple and easy interface: put in the z-value to get the probability area or the inverse. Even ask Siri!

R

R has functions `pnorm()` and `qnorm()`. If you have a Z value and want to find the area under the curve to the left of that value, use `pnorm(X)`. If you don't tell it otherwise, it assumes mean is zero and standard deviation is one. If you want other mean/stdev combinations, add those – so leaving them out is same as `pnorm(X, mean = 0, sd = 1)` or change `0` and `1` as you wish. If you have a probability and want to go backwards to find X, then use `qnorm(p)`.

> ***Side Note**: The basic property, that the distribution is normal whatever the time interval, is what makes the normal distribution {and related functions, called Lévy distributions} special. Most distributions would not have this property so daily changes could have different distributions than weekly, monthly, quarterly, yearly, or whatever!*

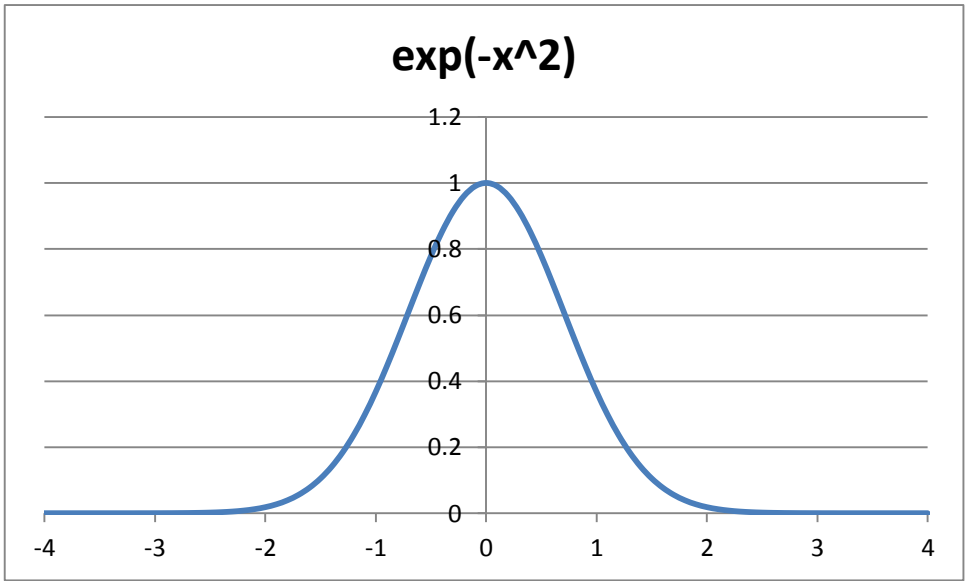
Recall from calculus the idea that some functions are not differentiable in places – they take a turn that is so sharp that, if we were to approximate the slope of the function coming at it from right or left, we would get very different answers. The function, $y = |x|$, is an example: at zero the left-hand derivative is -1; the right-hand derivative is 1. It is not differentiable at zero – it turns so sharply that it cannot be well approximated by local values. But it is continuous – it can be continuous even if it is not differentiable.

Now suppose I had a function that was everywhere continuous but nowhere differentiable – at every point it turns so sharply as to be unpredictable given past values. Various such functions have been derived by mathematicians, who call it a Wiener process; it generates Brownian motion. (When Einstein visited CCNY in 1905 he discussed his paper using Brownian motion to explain the movements of tiny particles in water, that are randomly bumped around by water molecules.) This function has many interesting properties – including an important link with the Normal distribution. The Normal distribution gives just the right degree of variation to allow continuity – other distributions would not be continuous or would have infinite variance.

Note also that a Wiener process has geometric form that is independent of scale or orientation – a Wiener process showing each day in the year cannot be distinguished from a Wiener process showing each minute in another time frame. As we noted above, price changes for any time interval are normal, whether the interval is minutely, daily, yearly, or whatever. These are fractals, curious beasts described by mathematicians such as Mandelbrot,

because normal variables added together are still normal. (You can read Mandelbrot's 1963 paper in the Journal of Business, which you can download from JStor – he argues that Wiener processes are unrealistic for modeling financial returns and proposes further generalizations.)

The Normal distribution has a pdf which looks ugly but isn't so bad once you break it down. It is proportional to e^{-x^2}. This is what gives it a bell shape:



To make this a real probability we need to have all of its area sum up to one, so the probability density function (PDF) for a standard normal (with zero mean and standard deviation of one) is

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

To allow a mean, μ, different from zero and a standard deviation, σ, different from one, we modify the formula to this:

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

The connection with *e* is useful if it reminds you of when you learned about "natural logarithms" and probably thought "what the heck is 'natural' about that ugly thing?!" But you learn that it comes up everywhere (think it's bad now? wait for differential equations!) and eventually make your peace with it. So too the 'normal' distribution.

If you think that the PDF is ugly then don't feel bad – its discoverer didn't like it either. Stigler's History of Statistics relates that Laplace first derived the function as the limit of a binomial distribution as $n \to \infty$ but couldn't believe that anything so ugly could be true. So he

put it away into a drawer until later when Gauss derived the same formula (from a different exercise) – which is why the Normal distribution is often referred to as "Gaussian". The Normal distribution arises in all sorts of other cases: solutions to partial differential equations; in physics Maxwell used it to describe the diffusion of gases or heat (again Brownian motion; video here http://fuckyeahfluiddynamics.tumblr.com/post/5678567551o/have-you-ever-noticed-how-motes-of-dust-seem-to); in information theory where it is connected to standard measures of entropy (Kullback Liebler); even in the distribution of prime factors in number theory, the Erdős–Kac Theorem.

Finally I'll note the statistical quincunx, which is a great word since it sounds naughty but is actually geeky (google it or I'll try to get an online version to play in class).