Lecture Notes

Econ B2000, Statistics and Introduction to Econometrics

Kevin R Foster, the Colin Powell School at the City College of New York, CUNY

Fall 2024 DRAFT

Why these lecture notes? What is the value added over the textbook? In my experience, econometrics textbooks are a great resource for "how" but not so useful for "why". The textbook tells you how to perform various analyses but the motivation is left exogenous. Of course for many students the motivation is simply to get a grade in a class, but I hope that I can convince you to be genuinely curious.

A textbook is usually structured in the way a brick wall is built: one layer gradually built up on another, with the base made solid before going on. My lectures on the other hand go in circles, making a quick dash into an advanced topic to pique your interest, then going back to fill in some of the basics, then dashing ahead again, sketching a link to another topic, generally just trying to be dynamic. I will leave it to you to fill in some of the holes, once I have convinced you that it's worthwhile. Learning has some aspects of prospect theory (which you should have done in micro theory) since prospect theory asks how people make rational decisions about completely unknown areas, trying to decide if it is worthwhile to invest in a blank spot – where one goal is to fill in the blanks. In this case, many students don't know much about how useful econometrics is so I want to persuade you, both in class and through these notes.

There is a reason that textbooks are this way: they try not to be wrong. A textbook is supposed to be scripture, giving you the capital-T Truth; this tends to make rather dull reading. These notes are more likely to be wrong. A famous statistician, Prof Box, said all models are wrong, but some are useful. So too with texts. Some of this material might be wrong, much more of it is certainly arguable. (As an example, statisticians hate the popular discussion of confidence intervals – but reading a true explanation is a real trial!) Sometimes learning is not so much acquiring the Truth as progressing through a series of approximations, each one closer and better. I hope you will become interested enough in the field to begin to argue and explore for yourself. Any text that gets a student interested must be doing something right. So please argue back at me.

Structure The first section (up to Discrete & Continuous Random Variables) provides background: lots of material that is important, that we'll use in class and in homework assignments and exams, that is fundamental to your progress in the course – but isn't that hard to learn. Parts may be a bit tedious but that's an occupational hazard. Some parts will be review and you should feel free to skip or skim those parts. The point is to get everybody up to a common level. Just don't skip the part on how to use R (unless you already know that). The rest of the sections should get about to the end of class – but note that I may be updating the post-midterm sections.

These notes are somewhat correlated with the Stock and Watson textbook but are not a substitute – read both.

Table of Contents

Beginning Notes

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.
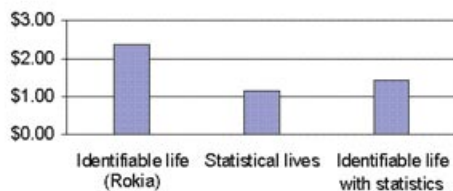
> A follow-up experiment by Small et al. initially primed study participants either to feel ("Describe your feelings when you hear the word 'baby,'" and similar items) or to do simple arithmetic calculations. Priming analytic thinking (calculation) reduced donations to the identifiable victim (Rokia) 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. We'll spend a lot of time getting on the computer and actually doing statistics.

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

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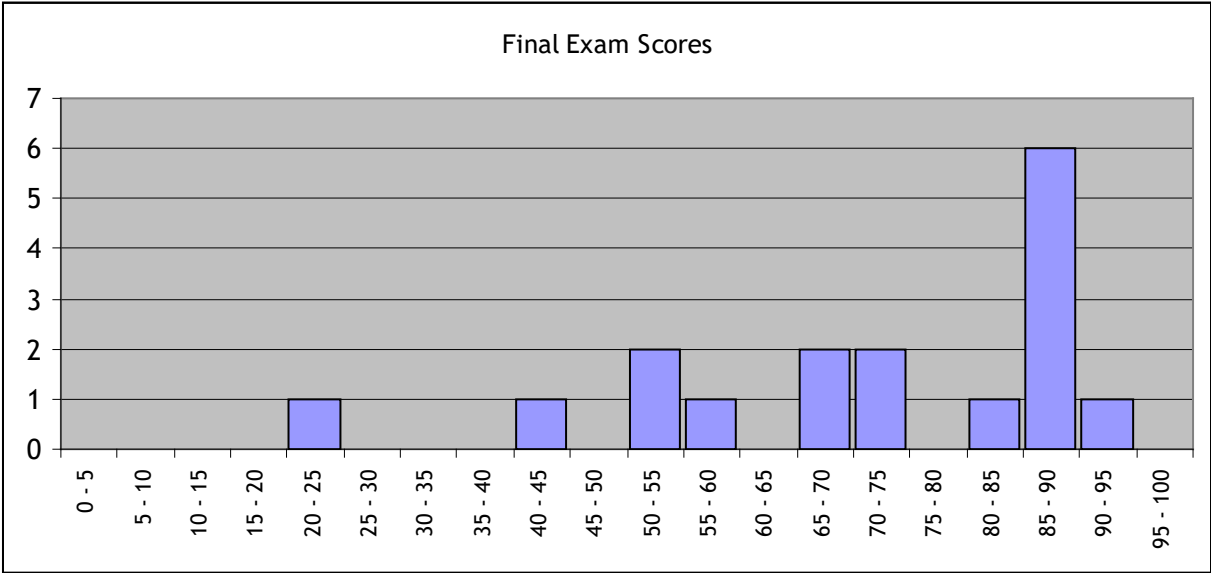
This histogram shows a great deal of information; more than just a single number could tell. (Although this histogram, with so many one- or two-step sizes, could be made much better.)

Often a histogram is presented, as above, with blocks but it can just as easily be connected lines, like this:
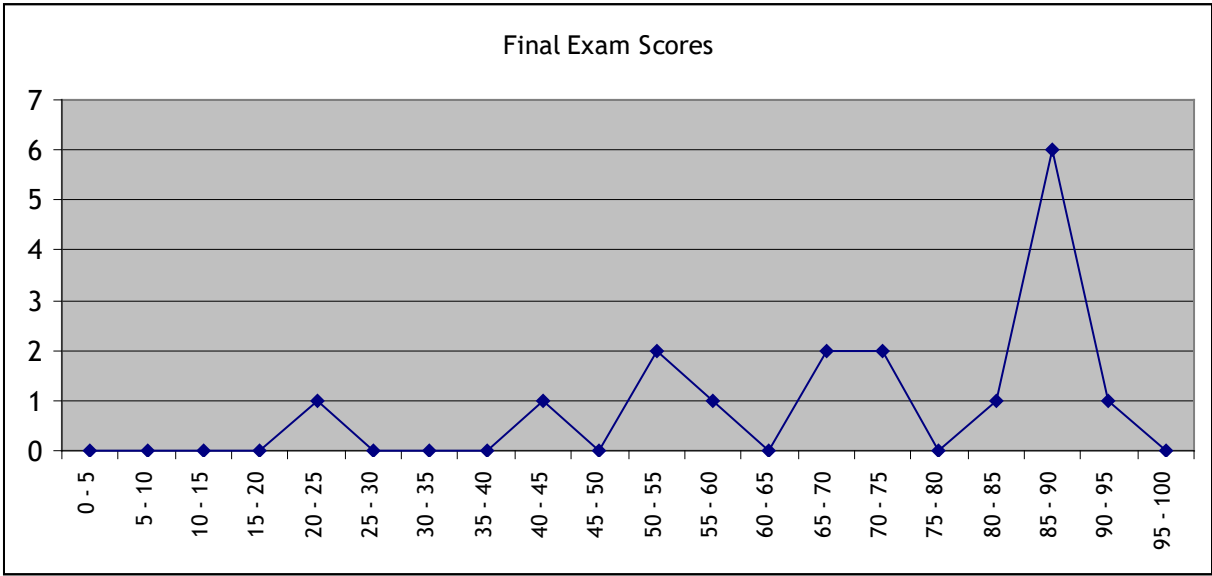
The information in the two charts is identical.

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:
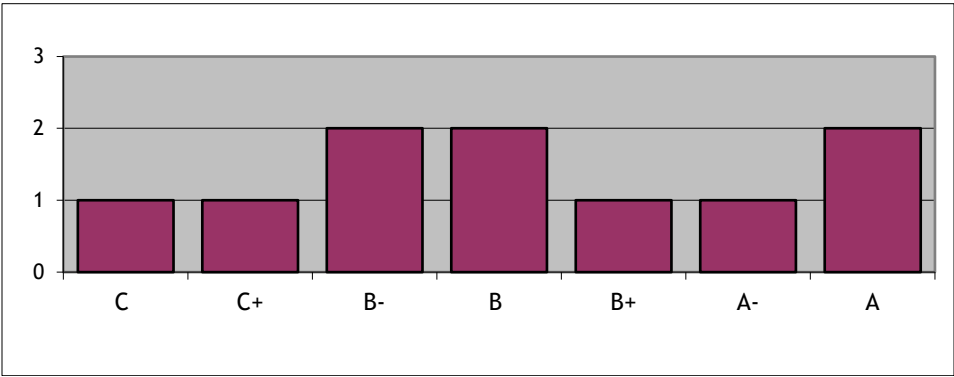
whereas by collapsing together the grades into A, B, and C categories we would get something more intelligible, like this:
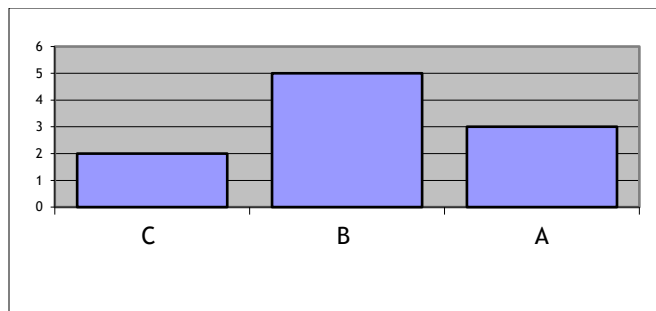
This shows the central tendency much better – the student has gotten many B grades and slightly more A grades than C grades. The previous histogram had too many categories so it was difficult to see a pattern.

Another reason to show the data is to reveal structure that simple averages wouldn't show. Consider the "datasaurus" where each scatter plot below has the same X and Y means, standard deviations, and correlation (by Alberto Cairo https://www.autodeskresearch.com/publications/samestats):

Fig 2. The **Datasaurus Dozen**. While different in appearance, each dataset has the same summary statistics (mean, standard
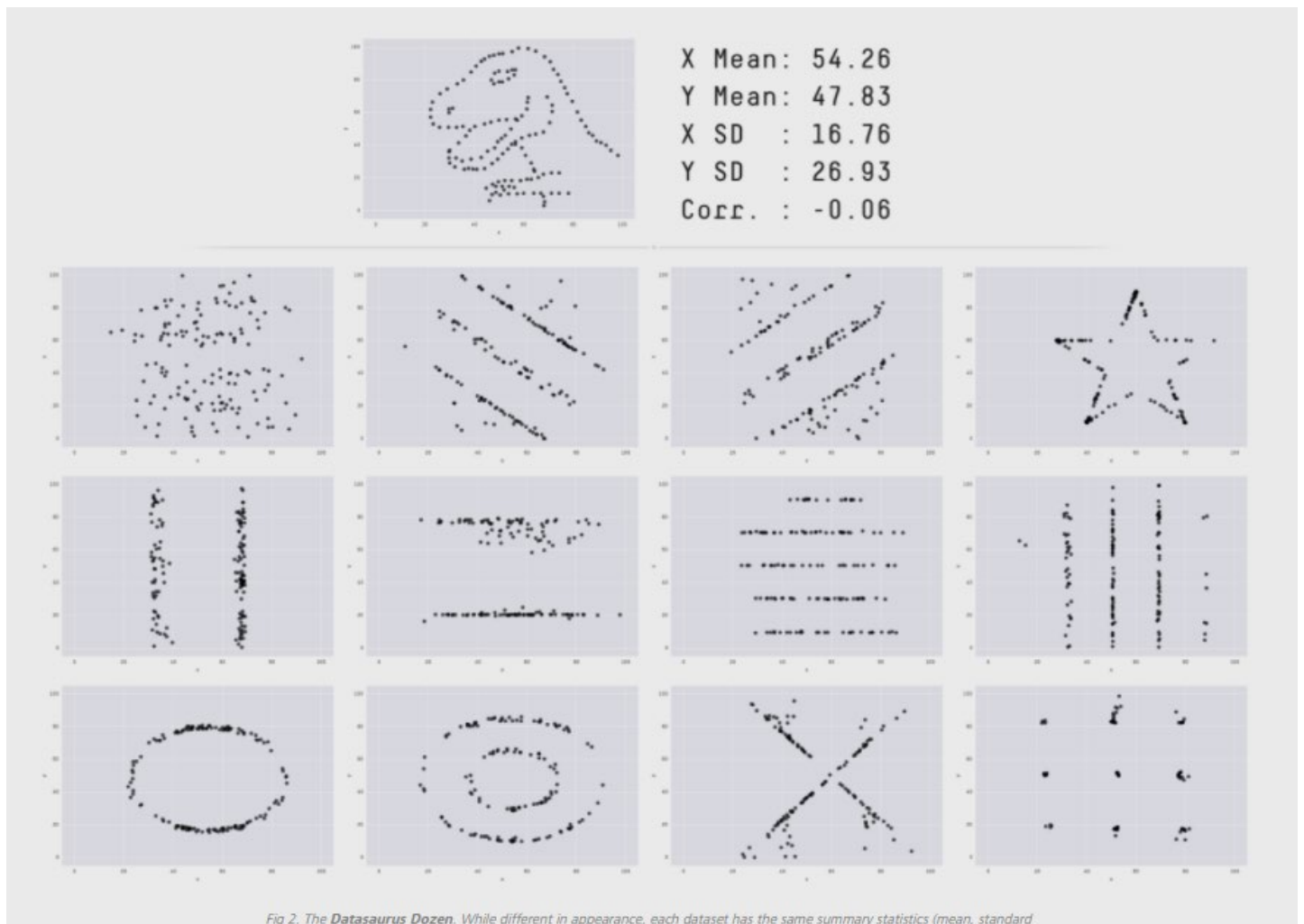
Basic Concepts: Find the Center of the Data

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

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

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):
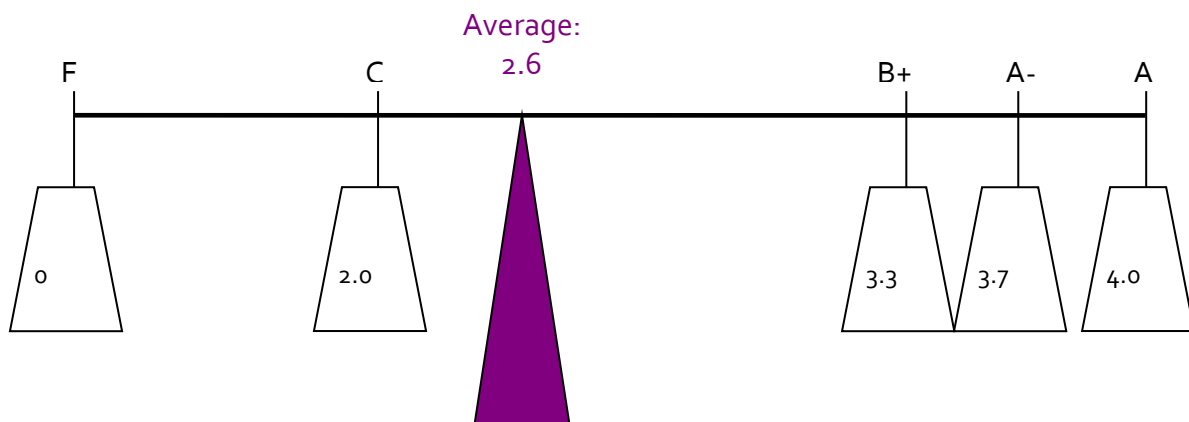
So the "bonus" example would look like this, with one person getting $100,000 while the other 99 get nothing:
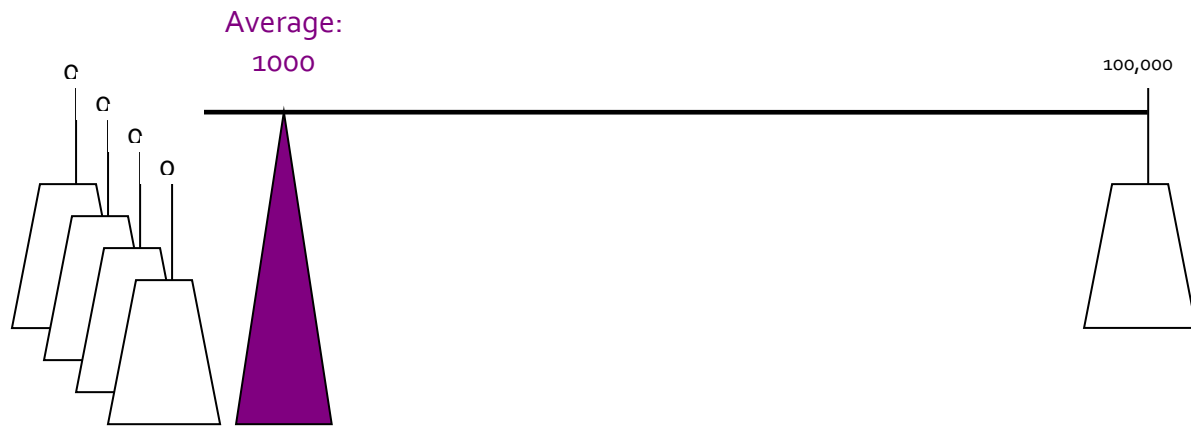
Where there are actually 99 weights at "zero." But even one person with such a long moment arm can still shift the center of gravity away.

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)

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 $\bar{X} = \dfrac{1}{N}\sum\limits_{i=1}^{N} X_i$.

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

We know that $\sum_{i=1}^{N}\left(X_i - \overline{X}\right) = \sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \overline{X}$ and, since $\overline{X}$ is the same for every observation,

$\sum_{i=1}^{N} \overline{X} = N\overline{X} = \sum_{i=1}^{N} X_i$, if we substitute back from the definition of $\overline{X}$. So $\sum_{i=1}^{N}\left(X_i - \overline{X}\right) = 0$. We can't re-use the

average. So we want to find some useful, sensible function [or functions], $f(\cdot)$, such that $\sum_{i=1}^{N} f\left(X_i - \overline{X}\right) \neq 0$.

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:

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

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

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

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

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

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

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 $\sum_{i=1}^{N} f(X_i - \bar{X}) \neq 0$ -- for example, the mean of the absolute value, $\frac{1}{N}\sum_{i=1}^{N}|X_i - \bar{X}|$. By recalling the graphs of these two functions you can begin to appreciate how they differ:
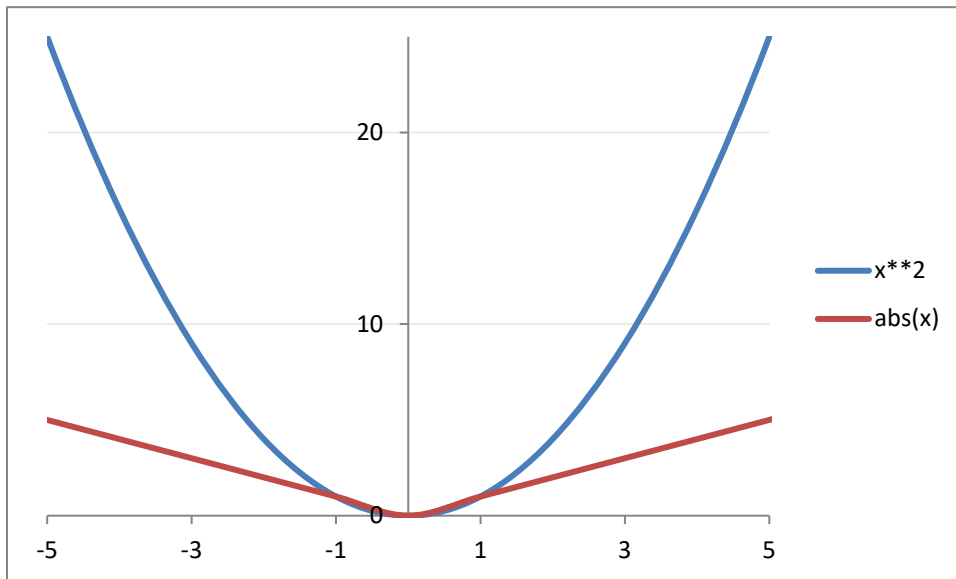
So that squaring the difference counts large deviations very much worse than small deviations, whereas an absolute deviation does not. So if you're trying to hit a central target, it might well make sense that wider and wider misses should be penalized worse, while tiny misses should be hardly counted.

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

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

Sometimes you will see other measures of variation; the textbook goes through these comprehensively. Note that the Coefficient of Variation, $\frac{s}{\overline{X}}$, is the reciprocal of the signal-to-noise ratio.

This is an important measure when there is no natural or physical measure, for example a Likert scale. If you ask people to rate beers on a scale of 1-10 and find that consumers prefer Stone's Ruination Ale to Budweiser by 2 points, you have no idea whether 2 is a big or a small difference – unless you know how much variation there was in the data (i.e. the standard deviation). On the other hand, if Ruination costs \$2 more than Bud, you can interpret that even without a standard deviation.

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

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

Now Do It!

We'll use data from a variety of sources 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. For now you can skim these short descriptions of some datasets.

Household Pulse

This data from Census Bureau was begun in pandemic to take a "pulse" of how people were making various decisions so it was updated every few weeks.

Overview of PUMS

At times in the class 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. I give an overview of the basics of how to use that program.

The dataset (just people in the state of New York) has information on almost 200,000 people in almost 100,000 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.). 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.

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 thousands of 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]
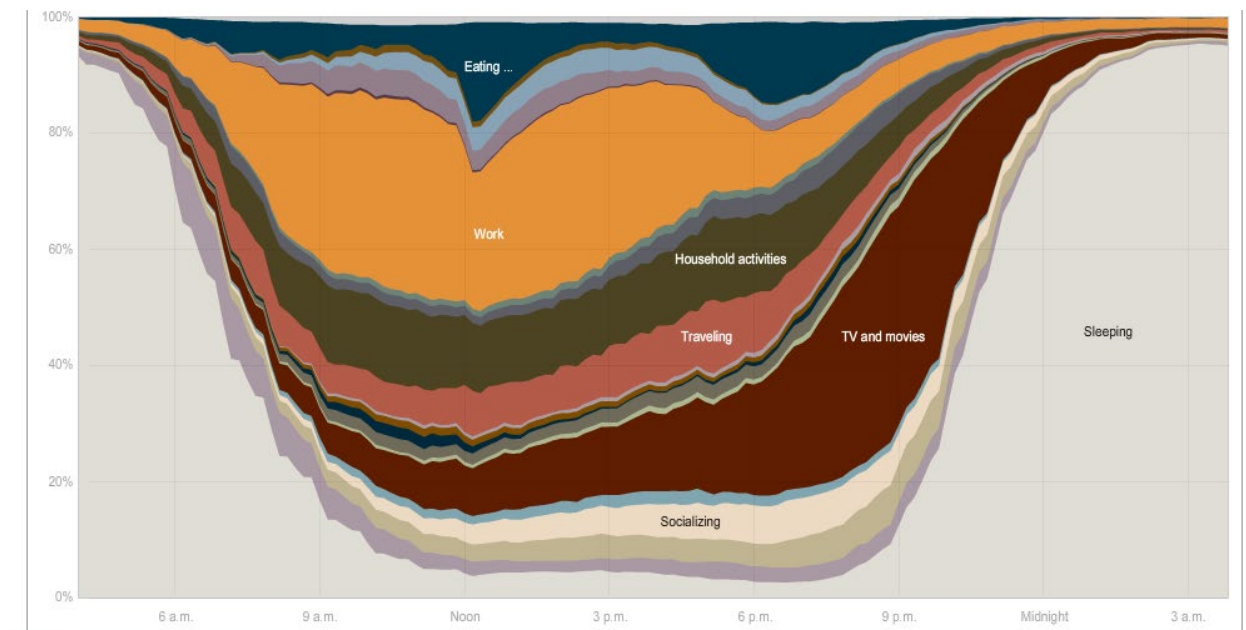
To use the data effectively, it is helpful to understand the ATUS classification system, where additional numbers at the right indicated additional specificity. The first two digits give generic broad categories. The general classification **T05** refers to time spent doing things related to work. **T0501** 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.

Consumer Expenditure Data

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

Taxi Data

The data is the "Fare Data" from andresmh.com/nyctaxitrips, which posts data originally from Chris Whong. *(Read the page about his FOIL request for the data, it's not often you find the phrase, "Overall, I have to say I was impressed with the TLC's responsiveness, professionalism, and the fact that they allow email correspondence for this sort of thing in the first place.")*

The TLC tried to make the medallion and hack licenses anonymous but messed up, so Vijay Pandurangan was able to actually decode, making it possible to figure out exactly what taxi went where and when.

I downloaded the first chunk (of 12) of the taxi data; it has 14,776,615 observations so working with it slows down my little laptop. But it gets the job done. If you have fun with it, grab the rest of the data and have more fun.

Note that this is not the full population of cab rides in 2013 but just a convenience sample (the first data chunk that was available) so it is NOT a random sample and so canNOT be interpreted as implying anything about the population. Nevertheless it's fun so for example you can figure out that tips are hugely under-reported, since only .01% of cash rides record any tip while 97% of credit-card rides report a tip. There might be occasions where the ride is paid with a card but the tip is in cash.

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

This survey is only made once every three years. 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 many millions, 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.

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 $(\bar{X}|Y=0), (\bar{X}|Y=1)$ *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)
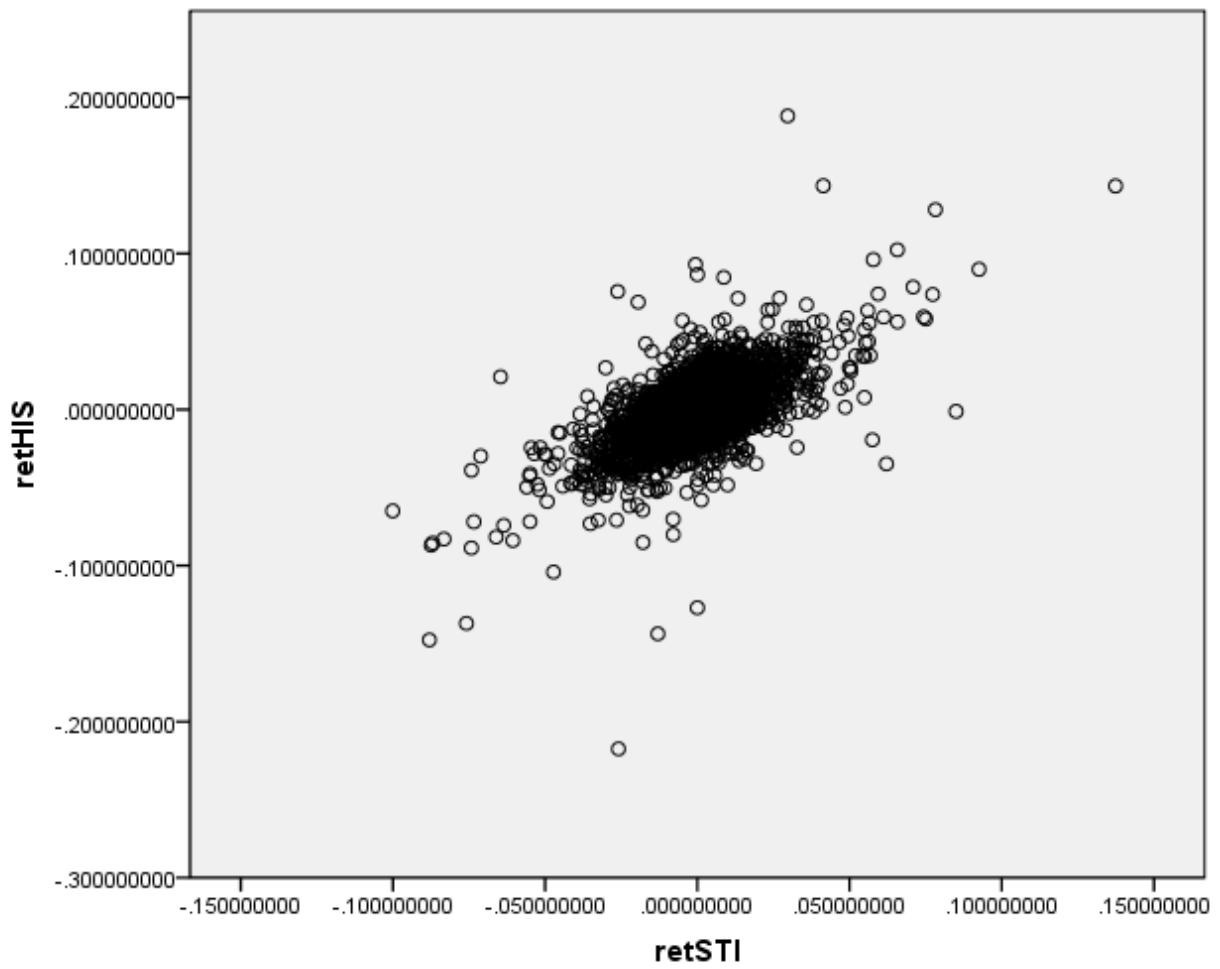
This next graph shows the S&P 500 returns and interest rates (1-month Eurodollar) during Jan 2, 1990 – Sept 1, 2010.
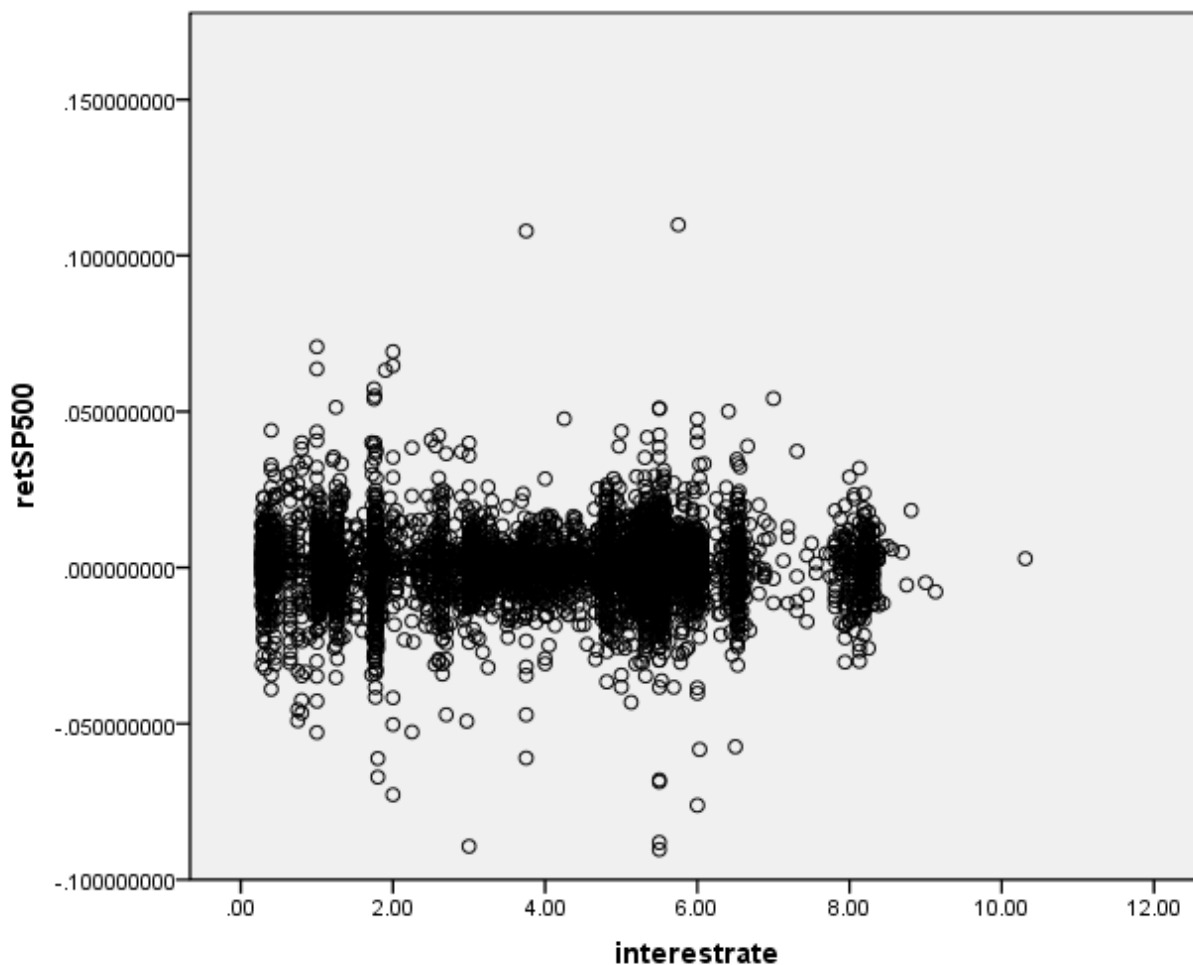
You don't have to be a highly-skilled econometrician to see the difference in the relationships. It would seem reasonable to state that the Hong Kong and Singapore stock returns are closely linked; while US stock returns are not closely related to US interest rates. (Remember, in most economic applications 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, $\overline{XY} = \dfrac{1}{n}\sum_{i=1}^{n} X_i Y_i$. 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:

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

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.

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 $Z_{X,i} = \dfrac{X_i - \bar{X}}{s_X}$, and the standardized value of Y is $Z_{Y,i} = \dfrac{Y_i - \bar{Y}}{s_Y}$.)

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, $\dfrac{1}{N}\sum_{i=1}^{N} Z_{X,i} Z_{Y,i}$.

This can be written as $\dfrac{1}{N}\sum_{i=1}^{N} Z_{X,i} Z_{Y,i} = \dfrac{1}{N}\sum_{i=1}^{N}\left(\dfrac{X_i - \bar{X}}{s_x}\right)\left(\dfrac{Y_i - \bar{Y}}{s_Y}\right) = \dfrac{1}{N}\dfrac{1}{s_x s_Y}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$. 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	σ11			

X2	σ21	σ22		
X3	σ31	σ32	σ33	
X4	σ41	σ42	σ34	σ44

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?

In any research question we will get a variety of answers and we try to separate out the signal from the noise. To answer, 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_o. 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 Games

As an example, consider games or 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 and basketball use 7 games to decide a champion; global football and American football use 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)	(L,L,W)	(L,L,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)	(1-p)(1-p)p	(1-p)(1-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...

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 – assume that 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 (we'll use the PUMS data on people in NY). 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 is:

	No HS	HS	SmColl	Bach	Adv
less than 10,000	15790	32307	16584	10490	7603
10,001 - 50,000	3484	16629	11205	8568	4118
50,001 - 100,000	494	5191	5089	7688	6571
100,001+	88	976	1134	4056	5093

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 marginals:

	No HS	HS	SmColl	Bach	Adv	Marginals
less than 10,000	0.097	0.198	0.102	0.064	0.047	0.507
10,001 - 50,000	0.021	0.102	0.069	0.053	0.025	0.270
50,001 - 100,000	0.003	0.032	0.031	0.047	0.040	0.153
100,001+	0.001	0.006	0.007	0.025	0.031	0.070
Marginals	0.122	0.338	0.208	0.189	0.143	

Some R code to do those tables:

```
table(cut(income_wagesal,breaks=4),educ_indx)
# but that output might not be quite what we want so instead tell it what
breaks to use
table(cut(income_wagesal,breaks=c(-
1000,10000,50000,100000,1000000)),educ_indx)
# note that I first used summary(income_wagesal) to figure max min and guess
at suitable break points
barplot(table(cut(income_wagesal,breaks=c(-
1000,10000,50000,100000,1000000)),educ_indx))
#alternately
plot(cut(income_wagesal,breaks=c(-
1000,10000,50000,100000,1000000)),educ_indx)
```

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	Bach	Adv
less than 10,000	0.795	0.586	0.488	0.341	0.325
10,001 - 50,000	0.175	0.302	0.329	0.278	0.176
50,001 - 100,000	0.025	0.094	0.150	0.250	0.281
100,001+	0.004	0.018	0.033	0.132	0.218

This shows that, of the people without a high school diploma, 79.5% have wage of $10,000 or less, while just 32.5% of people with an Advanced Degree make that little money. On the opposite end, just about 4/10 of 1% of people without a high school diploma make over $100k, while nearly 22% 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 in the original table by its sum:

Conditional on Wage	No HS	HS	SmColl	Bach	Adv
less than 10,000	0.191	0.390	0.200	0.127	0.092
10,001 - 50,000	0.079	0.378	0.255	0.195	0.094
50,001 - 100,000	0.020	0.207	0.203	0.307	0.262
100,001+	0.008	0.086	0.100	0.357	0.449

So this shows that, of people making more than $100,000 in wages, 45% of them have an Advanced Degree, another 36% have a Bachelor's Degree, while just 27% 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? Smells like ... HOMEWORK!)

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. (Occasionally politicians think it might be a good idea to have, for example, every welfare recipient tested for drugs, without discussion of how many false positives and false negatives would be produced.)

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 code 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 code, "|". 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)
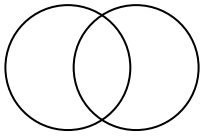
General Law of Addition

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

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

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

Counting Rules

If A can occur as N₁ events and B can be N₂ events then the sample space is $N_1 \cdot N_2$ (visualize a contingency table with N₁ rows and N₂ columns).

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

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 Distributions:

Uniform

- depend on only upper and lower bound, so all events are in $[a,b]$
- mean is $\dfrac{a+b}{2}$; standard deviation is $\sqrt{\dfrac{[b-a+1]^2 - 1}{12}}$

- 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.
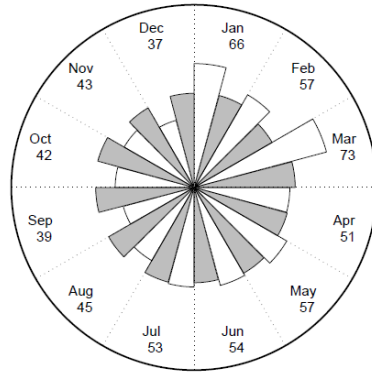

Figure 1: Circular plot of the observed and expected number of AFL players' births. The observed values are shown in white segments and the expected value in grey. The numbers around the outside of the plot are the observed number of births in each month. The expected number of births are based on national 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 R program. This is a warning that sometimes our intuition about how we might think numbers are distributed is actually wrong.

Bernoulli

- depend only on p, the probability of the event occurring

- calc mean and standard deviation from formulas, $E(X) = \mu = \sum_{i=1}^{N} x_i P(x_i)$ and

$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 P(x_i)$. Where there are only 2 values, 0 and 1, this is easy to calculate. E(X) here is 1*P(x=1)+0*P(x=0) = 1*p + 0*(1-p) = p. Variance is (1-p)² * p + (0 – p)² * (1-p) = {some algebra to write out} = p – p².

  - mean is p; standard deviation is $\sqrt{p(1-p)}$

    o   *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 intuition.*
  - 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, each independent; record how many were 1 not zero

- $\mu = np$; $\sigma = \sqrt{np(1-p)}$

o   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?

o   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 offered IPOs?

- questions about opinion polls – the famous "plus or minus 2 percentage points" – get margin of error depending on sample size (n)

Some students are a bit puzzled by two different sets of formulas for the binomial distribution – the standard deviation is listed as either $\sqrt{np(1-p)}$ or $\sqrt{\dfrac{p(1-p)}{n}}$. 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 $\sqrt{p(1-p)} = \sqrt{\dfrac{1}{2} \cdot \dfrac{1}{2}} = \sqrt{\dfrac{1}{4}} = \dfrac{1}{2}$. 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 $\sqrt{10} \cdot \dfrac{1}{2}$ 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 $\sqrt{500} \cdot \dfrac{1}{2}$ = 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 $\dfrac{\sqrt{n}}{n} = \dfrac{1}{\sqrt{n}}$. 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**.

Alternately instead of memorizing formulas for different distributions, you can derive this one easily from our rules of linear combinations. (Try it!)

Poisson

- model arrivals per time, assuming independent
- depends only on λ which is also mean
- PDF is $\dfrac{\lambda^x e^{-\lambda}}{x!}$
- 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:
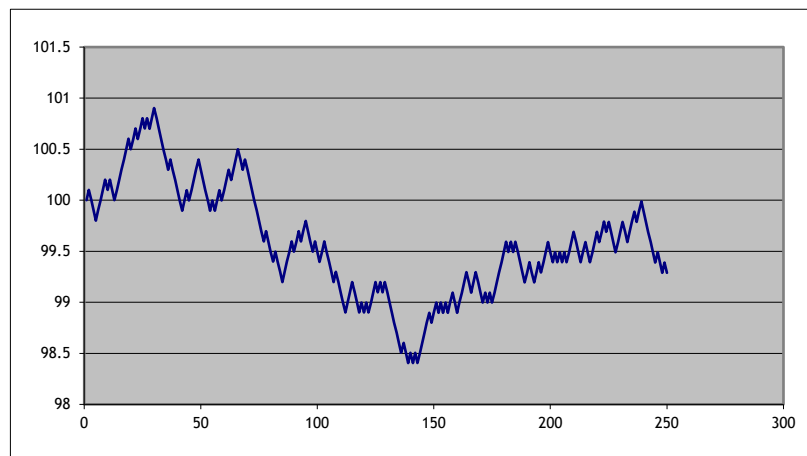
Here are 10 series (copied and pasted the whole S, "up," and "down" 10 times), see Excel sheet " *simple_stock_example_for_lecture2.xlsx*".
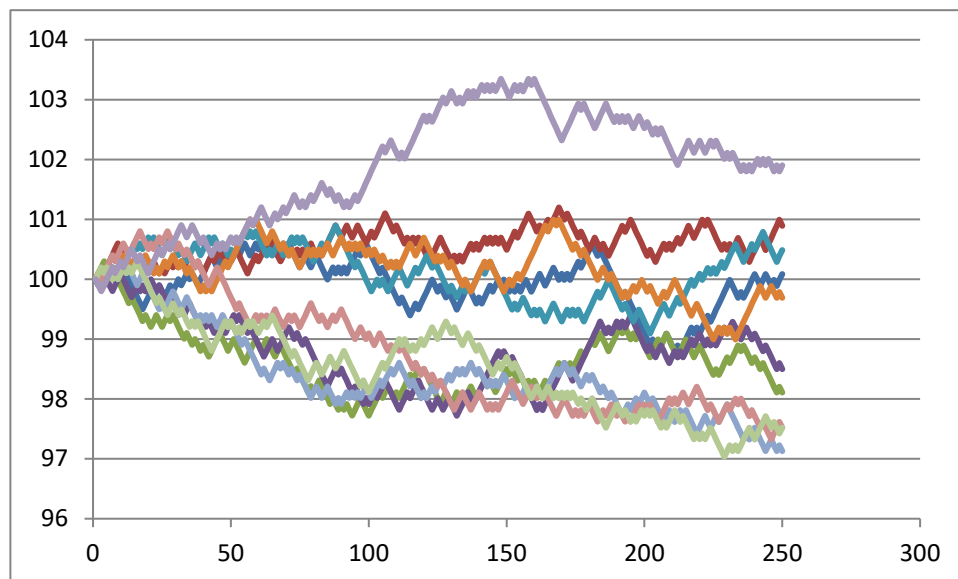
We're not done yet; we can make it better. But the real point for now is to see the basic principle of the thing: we can simulate stock price paths as random trips.

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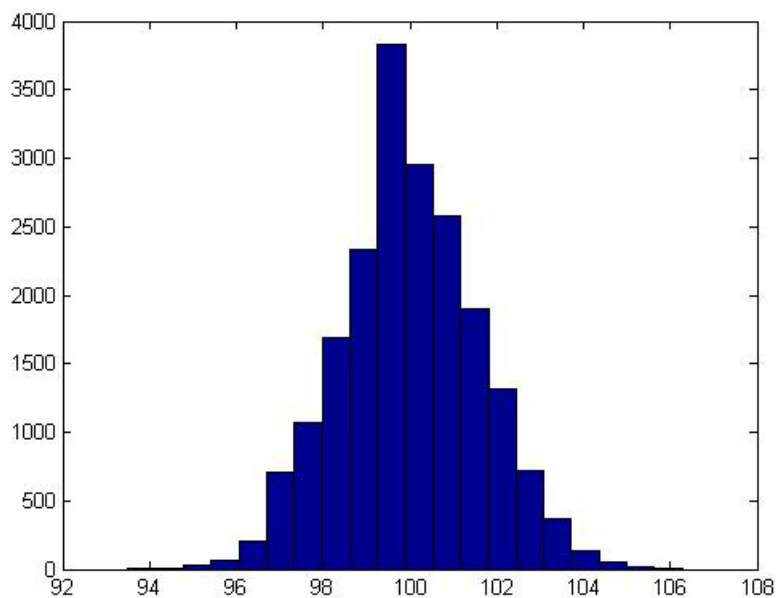
(If you're confident with your R knowledge, try writing that code!)

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. (See "Side Note" below for more detail.)

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:
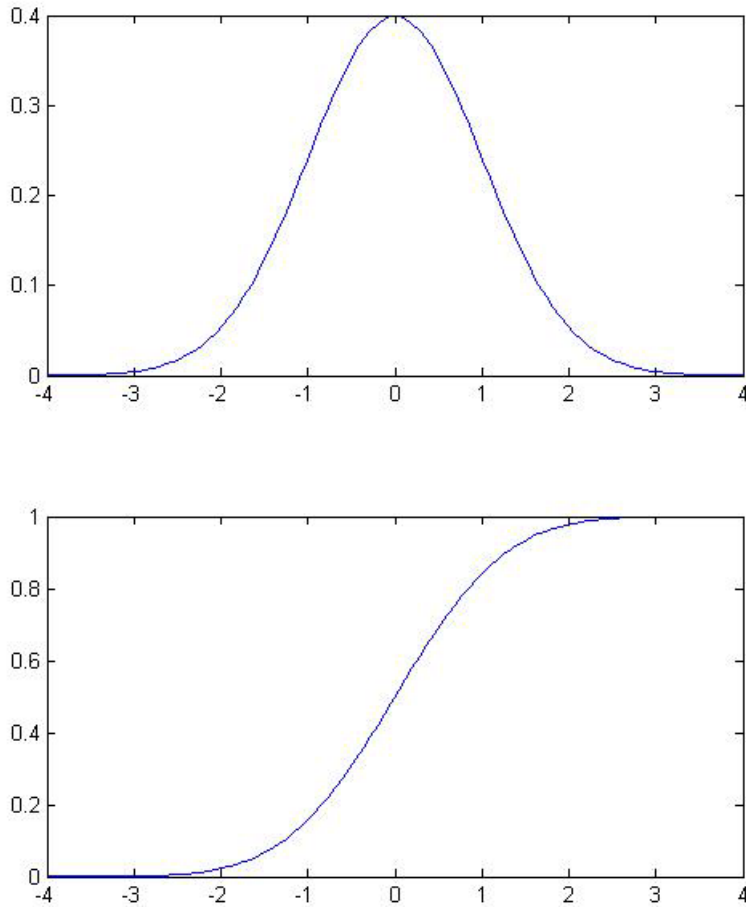
Example of using normal distributions:

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

A   NH Land, Jun–Jul–Aug

Normal Distribution
1951–1961
1961–1971
1971–1981
1981–1991
1991–2001
2001–2011
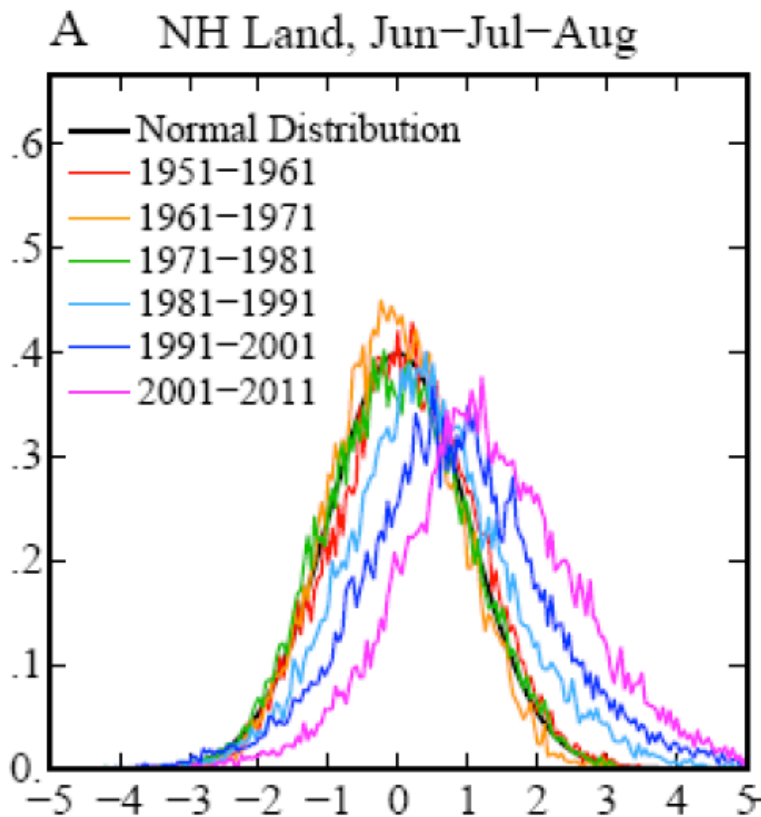
This shows the rightward spread of temperature deviations. The x-axis is in standard deviations, which makes the various geographies easily comparable (a hot day in Alaska is different from a hot day in Oklahoma). The authors define extreme heat as more than 3 standard deviations above the mean and note 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 + bμ) and standard deviation bσ. 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.

In some machine learning situations, data might be standardized, i.e. subtract the mean and divide by standard deviation, so $Z = \frac{X - \bar{X}}{s_X}$; or scaled to unit interval, so $W = \frac{X - X_{min}}{(X_{max} - X_{min})}$. Since these are linear transformations, we understand how these affect the distributions.

Oppositely, if we are creating random variables with a normal distribution, 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.

In R, just use rnorm() to get random numbers from a normal distribution; you can multiply and add to get other mean/stdev or you can used the canned procedure, rnorm(n, mean = 0, sd = 1).

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.

For class we're using the PUMS NY data with 196,585 observations and for now concentrate on the income from wages (INCWAGE) data. The true average of all of those people (omitting the na values) is $33,795.55. (Not quite; the top income value is cut at $638,000 – 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 data is 66,170.

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 $638,000!

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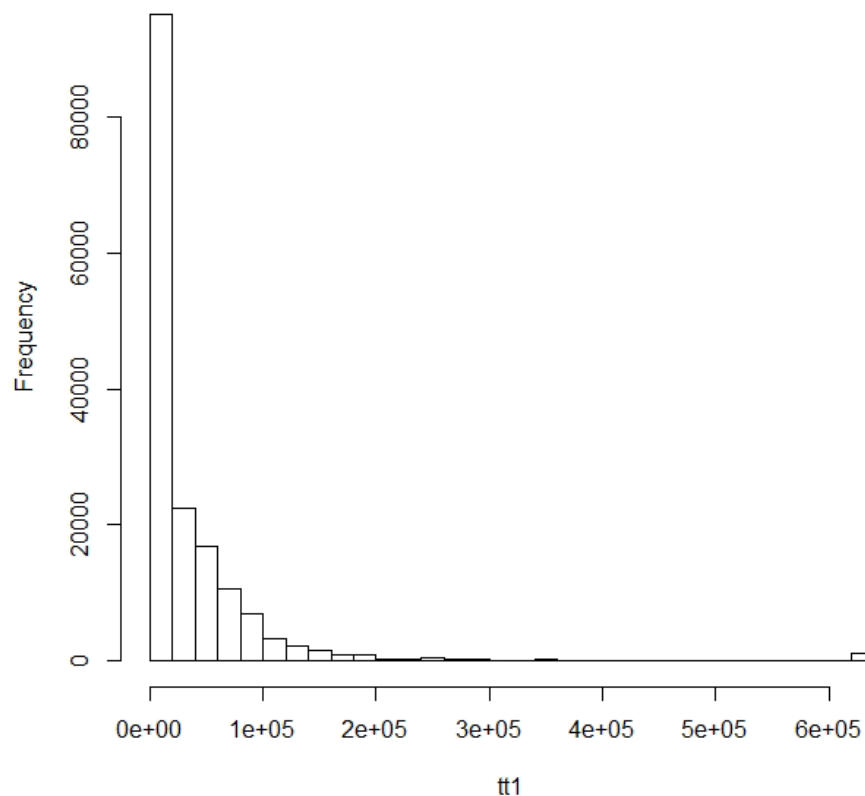
Taking an average of a population with such extreme values would seem to be difficult.

Suppose that I didn't want to calculate an average for all 196,585 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 10000 different polls (each poll with just 100 people).
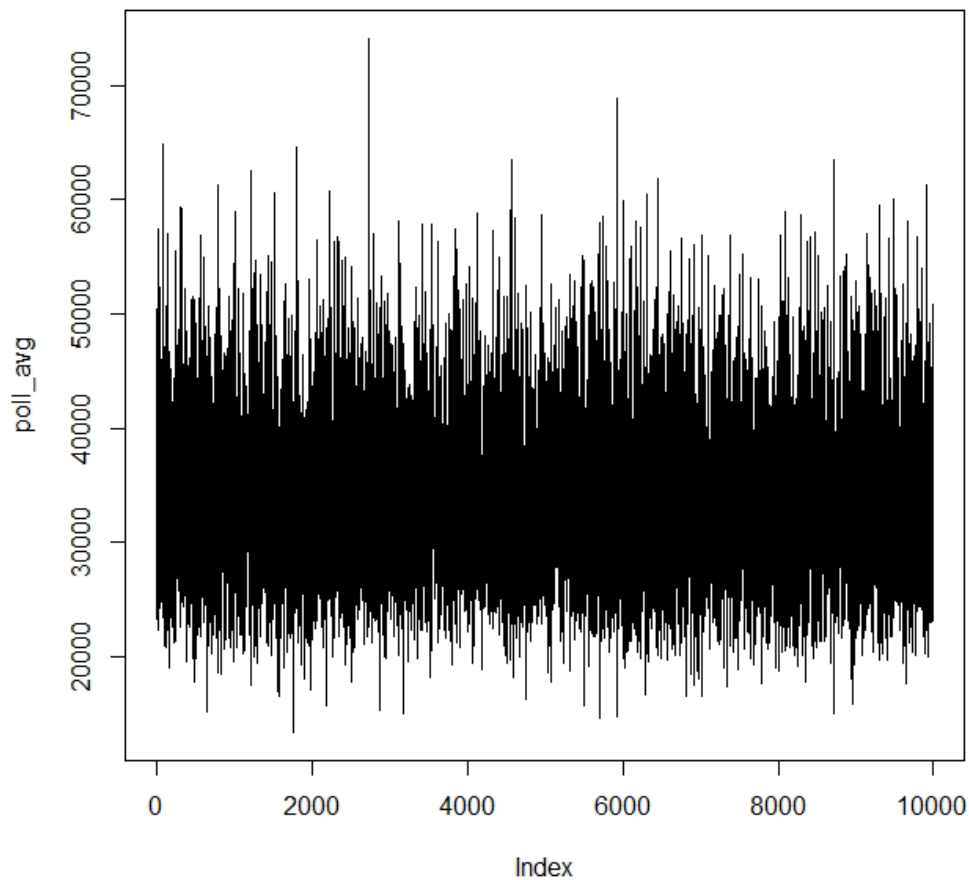
We can see that, although there are a few polls with averages as low almost 10,000 and a few with averages as high as 60,000, most of the polls are close to the true mean of $33,796.

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 10000 poll means is here:

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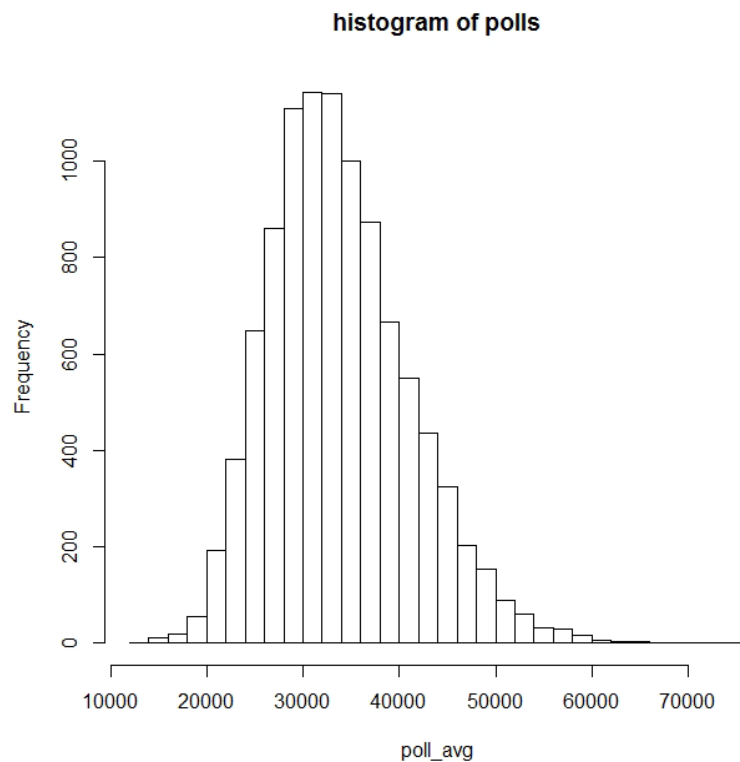
This shows that the distribution of the sample means looks like a Normal distribution – another case of how "normal" and ordinary the Normal distribution is.

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.
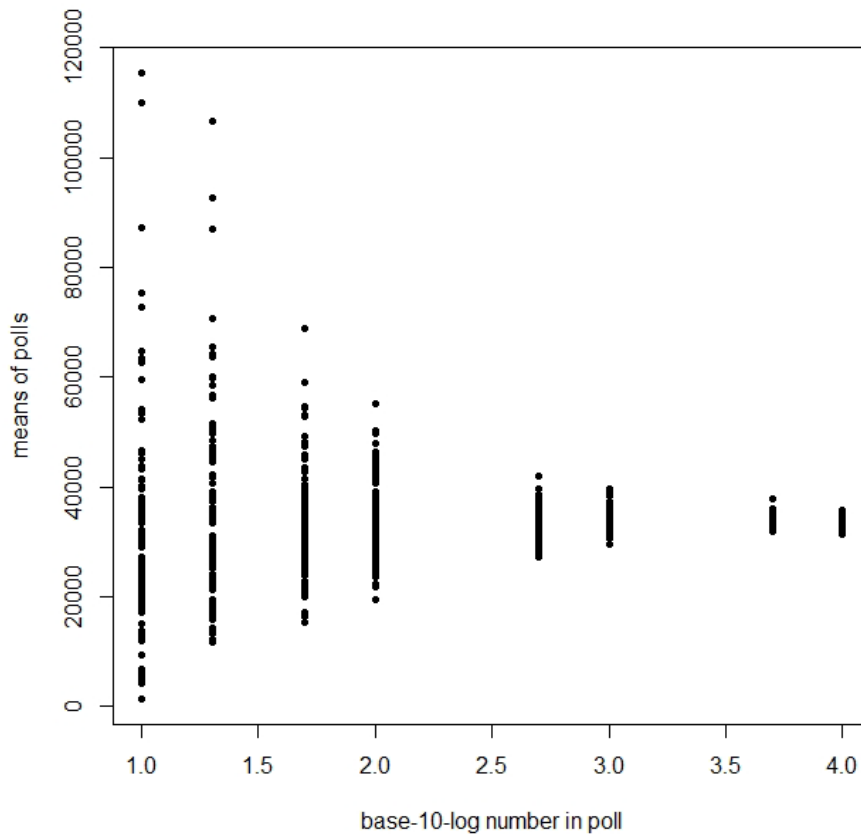
In the first set of 100 polls, on the left, each poll has just 10 people in it, so the results are quite varied. The next set has 20 people in each poll, so the results are closer to the true mean. By the time we get to 100 people in each poll (10^2 on the log-10-scale x-axis), the variation in the polls is much smaller. (Note that if you used the formulas from above instead of this Monte Carlo procedure, you would miss the asymmetry for the small polls.) As economists we would immediately see that there are diminishing marginal returns to sample size (and much of the business of polling derives from that).

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:
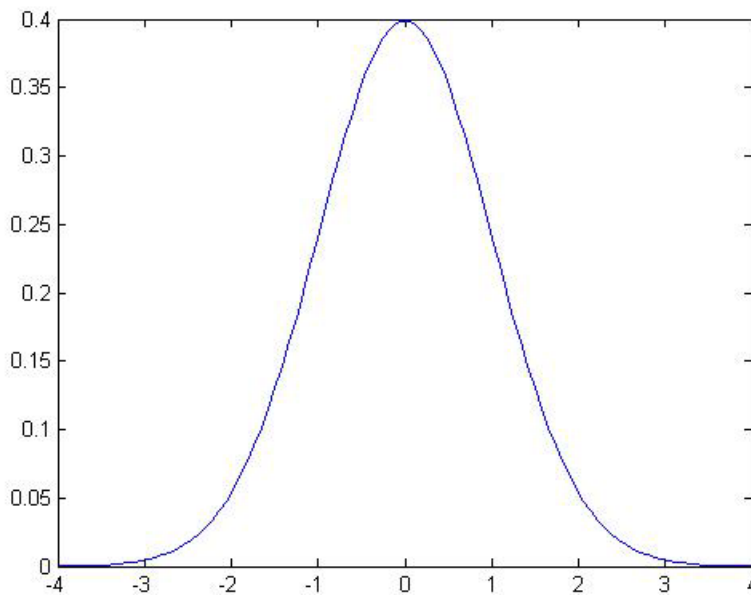
Note that it is symmetric around zero. Like any histogram, the area beneath the curve is a measure of the probability. The total area under the curve is exactly 1 (probabilities must add up to 100%). We can use the known function to calculate that the area under the curve, from -1 to 1, is 68.2689%. 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:
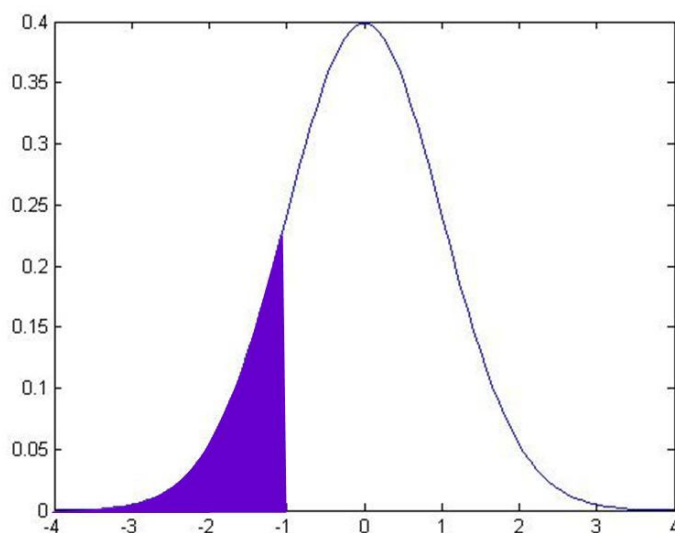
This area is 15.87%. How can I use this information to get the value that I earlier told you, that the area in between -1 and 1 is 68.2689%? Well, we know two other things (more precisely, I know them and I wrote them just 3 paragraphs up, so you ought to know them). We know that the total area under 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%.
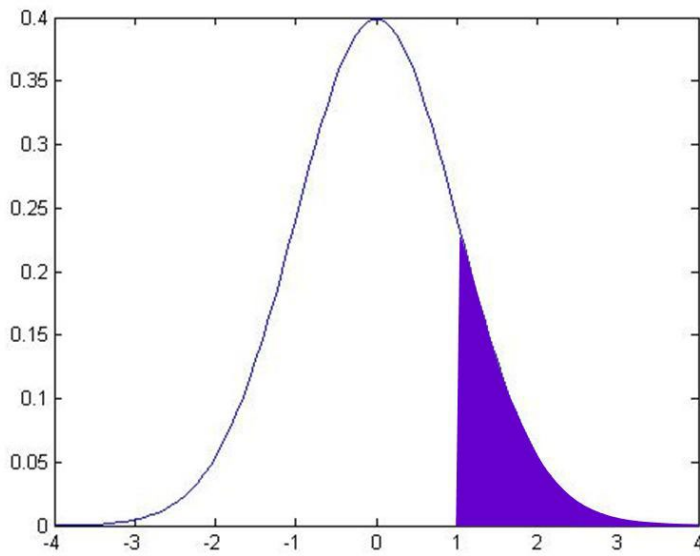
So to find the area in between -1 and +1, I take 100% and subtract off the two tail areas:
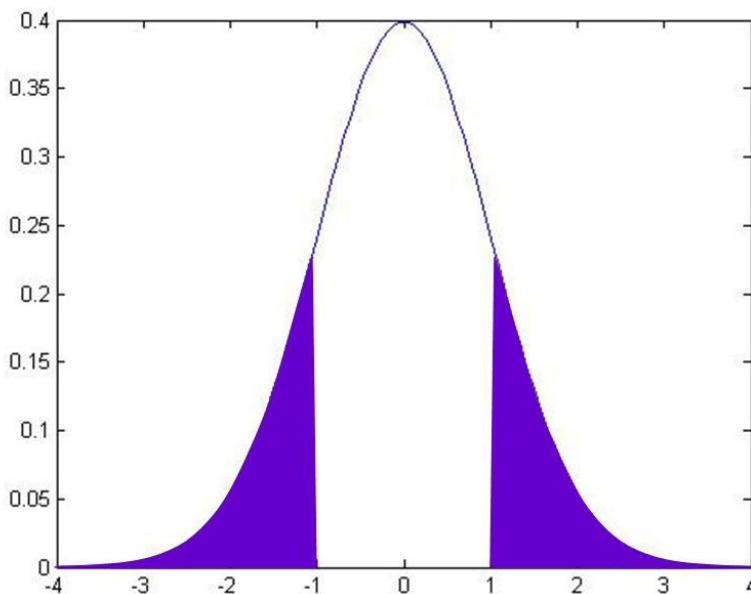
And this middle area is 100 − 15.87 − 15.87 = 68.26.

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, from negative infinity

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

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 `norm.s.dist` that assumes the mean is zero and standard deviation is one so you just use `norm.s.dist(X, TRUE)`. 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 `norm.s.inv`.

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 has a formula that 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:

exp(-x^2)
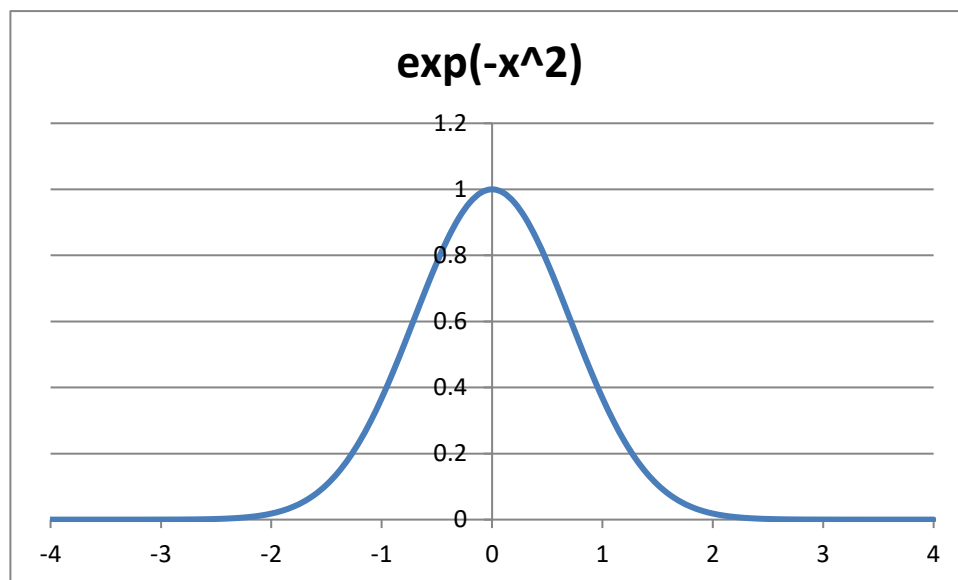
To make this a real probability we need to have all of its area sum up to one, so the probability density function (PDF) for a standard normal (with zero mean and standard deviation of one) is

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

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/56785675510/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.

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).

Final note on stratified sampling: Look again at this picture,
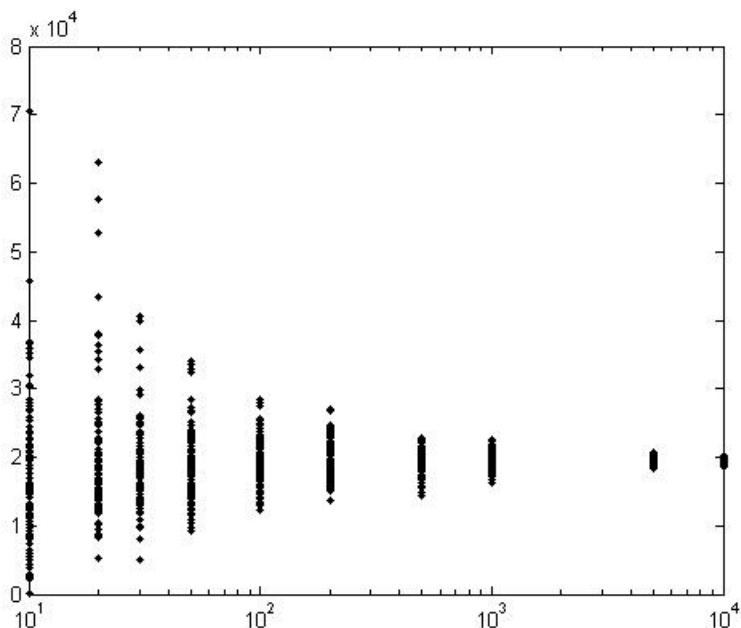
You can see, from the perspective of an economist, that the "production function" of accuracy as a function of the number of observations has diminishing returns – doubling the number of observations has a progressively smaller impact on accuracy. This is why many government data sets have weights for over-sampling of smaller populations. Suppose there are two groups of people; one makes up 90% of the population. Then if we randomly sample from the population, a sample of 1000 people would be expected to have 900 from one group (getting quite small standard errors) while just 100 from the other group (larger standard errors). The marginal increase in accuracy, from increasing the sample size, is very far from equal in the two groups. So many datasets oversample smaller populations – the equivalent of sampling 800 from the big group and 200 from the small group, then using the weights to fix the fact that the smaller group is oversampled. The exact procedures of weighting vary with the dataset. For this class, we will ignore the problem and not worry about the weights, but if you go on to do more stats, you can figure it out.

Is That Big?

Learning Outcomes (from CFA exam Study Session 3, Quantitative Methods)

Students will be able to:

- define the standard normal distribution, explain how to standardize a random variable, and calculate and interpret probabilities using the standard normal distribution;

The sample average has a normal distribution. This is hugely important for two reasons: one, it allows us to estimate a parameter, and two, because it allows us to start to get a handle on the world and how we might be fooled.

You calculate some statistic, maybe it's a difference between means of two groups. But you immediately have to answer: is that big? Is it a big difference?
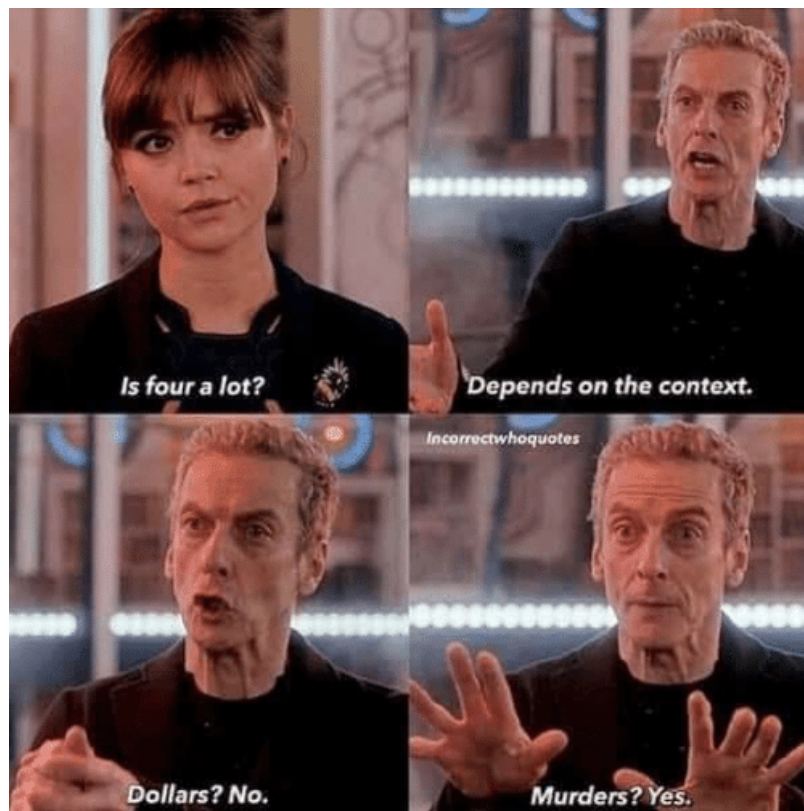
Well, it's about the standard error... We have to understand the issues of sampling.

Get a central parameter

The basic idea is that if we take the average of some sample of data, this average should be a good estimate of the true mean. For many beginning students this idea is so basic and obvious that you never think about when it is a reasonable assumption and when it might not be. For example, one of the causes of the Financial Crisis was that many of the 'quants' (the quantitative modelers) used overly-optimistic models that didn't seriously take account of the fact that financial prices can change dramatically. Most financial returns are not normally distributed! But we'll get more into that later; for now just remember this assumption. Later we'll talk about things like bias and consistency.

Return to the example of loading the dice, that we tried in the first homework assignment. Suppose we rolled 2 dice, and want to distinguish if either one is loaded. Call them "A" and "B". These are the results:

	A	B
Number of times roll 1 ⚀	4	2
... 2 ⚁	2	2
... 3 ⚂	5	4
... 4 ⚃	1	2

... 5 ⚄	4	4	
... 6 ⚅	4	6	

You might guess that B is loaded. But how likely is it? Could they both be fair?

A comes up with a 6 on 4/20 = 0.2; B comes up as 6 on 6/20 = 0.3. Both are higher than the expected value of 0.167. They are different but is that a big difference? (How big is 'big'?)

Thinking about Sampling Distributions

We tried to load dice, to get them to come up 6 more often. Suppose we want to test a dice to see if it actually comes up 6 more often, we could roll it once. If it comes up 6 then does that prove it's loaded? Well we know that a 6 comes up 1/6 of the time even with a fair dice, so that's not too improbable. What about if the first 2 rolls come up 6 – how likely is that, if the dice were fair? Well the likelihood of getting 2 rolls of 6 is (1/6)*(1/6) = 1/36, so that gets less likely, under 3%. The likelihood of getting a 6 three times in a row is even less, 1/6^3 = 1/216 = .0046. So if we keep rolling and keep getting a 6 each and every time, the likelihood of the dice being fair just keeps falling and falling. At some point we would decide that the likelihood of the dice being fair is just too low, and end the experiment.

But what if the dice came up 6 twice out of the first 3 rolls – would that be the same level of evidence? Again we might want to figure out how likely it would be, for a fair dice to come up with 2/3 of the rolls as a 6. This is a bit more of a complicated permutation since either the first, second, or third roll could be the non-6 roll. Recall that we can represent it (as if in extended form of game) as:

etc....
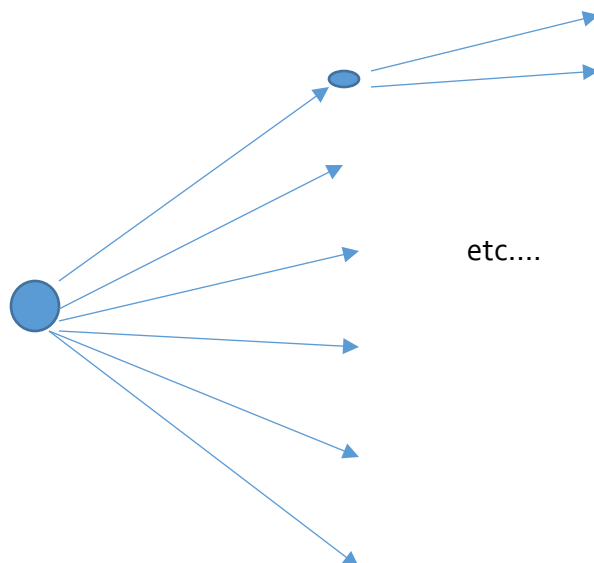
But quickly I get lazy and don't want to draw 6 choices, each with 6 choices, each with 6 choices, but instead represent the choice of rolling either a 6 or not-a-6, so

Roll a #

Roll a 6

#

6

#

6

#,#,#
#,#,6
#,6,#
#,6,6

6,#,#
6,#,6
6,6,#
6,6,6
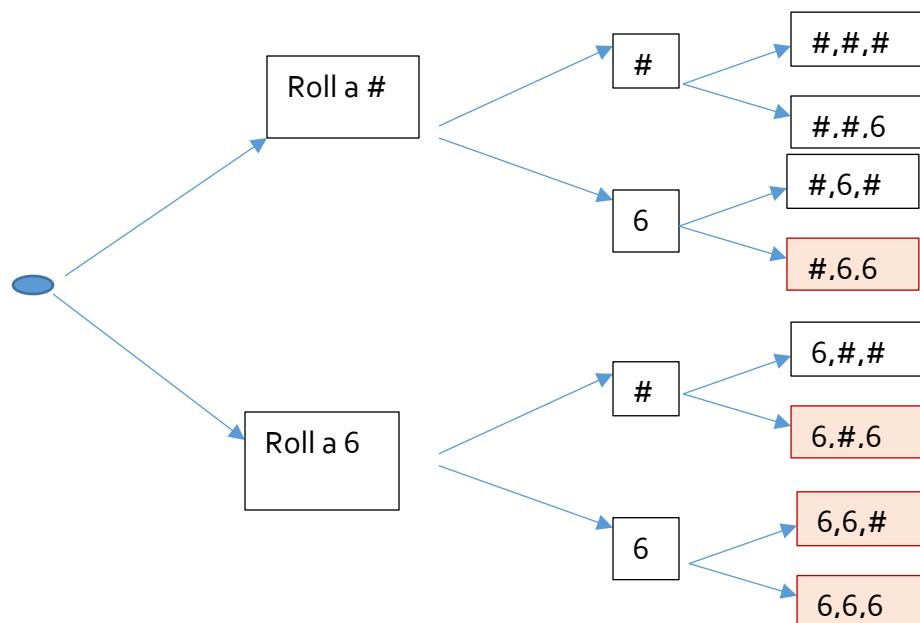
Then figure the probabilities of each outcome, where probability of rolling 6 is 1/6 and probability of rolling another number is 5/6.

Now I don't know about you, but I don't have the patience to do that for too many more rounds. If I roll the dice 10 times and want to see how likely it is, that at least 3 of the 10 rolls will come up 6 … that's just too much!

Fortunately we have a tool that is optimized for repeatedly doing very simple math problems, the computer. So fire up R!

```
# do one set of 10 rolls:
set.seed(12345)
x <- sample(6,10, replace = TRUE)
sum(x == 6)

# ----------------
NN = 100000
num_in_sampl <- rep(0,NN)
set.seed(12345)
for(indx in 1:NN) {
   x <- sample(6,10, replace = TRUE)
   num_in_sampl[indx] <- sum(x == 6)
}

h_s <- hist(num_in_sampl, breaks = c(-1,0,1,2,3,4,5,6,7,8,9,10))
prop.table(h_s$counts)
```

The next step is to ask, "do I have to do thousands of simulations every time?" Answer: "No, that's the power of stats!" Rather than doing a lot of simulations you can just find a formula. Sure the formula is a bit ugly but you've seen the program, it's not so easy either. (As you get more sophisticated you will find that there are tradeoffs to each method.)

Knowing that the sample average has a normal distribution also helps us specify the variation involved in the estimation. We often want to look at the difference between two sample averages, since this allows us to tell if there is a useful categorization to be made: are there really two separate groups? Or do they just happen to look different?

How can we try to guard against seeing relationships where, in fact, none actually exist?

To answer this question we must think like statisticians. To "think like a statistician" is to do mental handstands; it often seems like looking at the world upside-down. But as you get used to it, you'll discover how valuable it is. (There is another related question: "What if there really is a relationship but we don't find evidence in the sample?" We'll get to that.)

The first step in "thinking like a statistician" is to ask, What if there were actually no relationship; zero difference? What would we see? A big difference would be evidence in favor of different means; a small difference would be evidence against. But, in the phrase of Dierdre McCloskey, "How big is big?"

Law of Large Numbers

Probability and Statistics have many complications with twists and turns, but it all comes down to just a couple of simple ideas. These simple ideas are not necessarily intuitive – they're not the sort of things that might, at first, seem obvious. But as you get used to them, they'll become your friend.

One basic idea of statistics is the "Law of Large Numbers" (LLN). The LLN tells us that certain statistics (like the average) will very quickly get very close to the true value, as the size of the random sample increases. This means that if I want to know, say, the fraction of people who are right-handed or left-handed, or the fraction of people who will vote for Politician X versus Y, I don't need to talk with every person in the population.

This is strenuously counter-intuitive. You often hear people complain, "How can the pollsters claim to know so much about voting? They never talked to me!" But they don't have to talk to everyone; they don't even have to talk with very many people. The average of a random sample will "converge" to the true value in the population, as long as a few simple assumptions are satisfied.

With computers we can take much of the complicated formulas and derivations and just do simple experiments. Of course an experiment cannot replace a formal proof, but for the purposes of this course you don't need to worry about a formal proof.

R makes this easy. Run this little program, kind of like the dice example but for polling now:

```
# create the population of people
set.seed(1)
prob_of_yes <- 0.45
population_values <- runif(1000)
pop_yes <- (population_values < prob_of_yes)

# check that value should be near 0.45 although not exactly
mean(pop_yes)

# now do this the long way, for a sample of size 30 from the population
sampl_size <- 30
s1 <- sample(pop_yes,sampl_size)
mean(s1)
```

```
# you could go through and create s2, s3, etc or get lazy and do this...

# number of times to do this
NN <- 100

samples_from_pop <- matrix(data = NA, nrow = 1, ncol = NN)
for (i in 1:NN){
  samples_from_pop[i] <- mean(sample(pop_yes,30))
}
hist(samples_from_pop)

# you can go through and play with sample size, population size, and how many different
samples to take (NN)
```

You could do this with a spreadsheet, lots of formulas like "=if(RAND()<0.45,1,0)" but that's ugly! And it doesn't make it easy to replicate, but with "set.seed" you should be able to replicate the same results each time on R. (Read R's help on random numbers if you want to learn about pseudo-random number generation.)

In the problem set, you will be asked to do some similar calculations.

So we can formulate many different sorts of questions once we have this figured out.

First the question of polls: if we poll 500 people to figure out if they approve or disapprove of the President, what will be the standard error?

Standard Error of Average

With some math (⚡) we can figure out a formula for the standard error of the sample average. It is just the standard deviation of the sample divided by the square root of the sample size. So the sample average is distributed normally with mean of μ and standard error of $se = \frac{s}{\sqrt{N}}$. This is sometimes written compactly as $\overline{X} \sim N\left(\mu, \frac{s}{\sqrt{N}}\right)$.

Sometimes this causes confusion because in calculating the standard error, s, we divided by the square root of (N-1), since $s = \sqrt{\frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}{N-1}}$, so it seems you're dividing twice. But this is correct: the first division gets us an estimate of the sample's standard deviation; the second division by the square root of N gets us the estimate of the sample average's standard error.

The standardized test statistic (sometimes called Z-score since Z will have a standard normal distribution) is the mean divided by its standard error, $\frac{\overline{X}}{se} = \frac{\overline{X}}{s/\sqrt{N}} = \sqrt{N}\frac{\overline{X}}{s}$. This shows clearly that a larger sample size (bigger N) amplifies differences of $\overline{X}$ from zero (the usual null hypothesis). A small difference, with only a few observations, could be just chance; a small difference, sustained over many observations, is less likely to be just chance.

One of the first things to note about this formula is that, as N rises (as the sample gets larger) the standard error gets smaller – the estimator gets more precise. So if N could rise towards infinity then the

sample average would converge to the true mean; we write this as $\overline{X} \xrightarrow{p} \mu$ where the $\xrightarrow{p}$ means "converges in probability as N goes toward infinity".

So the sample average is **unbiased**. This simply means that it gets closer and closer to the true value as we get more observations. Generally "unbiased" is a good thing, although later we'll discuss tradeoffs between bias and variance.

Return to the binomial distribution, and its normal approximation. We know that std error has its maximum when p= ½, so if we put in p=0.5 then the standard error of a poll is, at worst, $\dfrac{1}{2\sqrt{n}}$, so more observations give a better approximation. See Excel sheet *poll_examples*. We'll return to this once we learn a bit more about the standard error of means.

⚡ A bit of Math:

We want to use our basic knowledge of linear combinations of normally-distributed variables to show that, if a random variable, X, comes from a normal distribution then its average will have a normal distribution with the same mean and the standard deviation of the sample divided by the square root of the sample size,

$$\overline{X} \sim N\left(\mu, \frac{s}{\sqrt{N}}\right).$$

The formula for the average is $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$. Consider first a case where there are just 2 observations. This case looks very similar to our rule about, if $W = CX + DY$, then $W \sim N\left(C\mu_X + D\mu_Y, \sqrt{C^2\sigma_X^2 + D^2\sigma_Y^2 + 2CD\sigma_{XY}}\right)$. With N=2, this is $\overline{X} = \dfrac{1}{2}X_1 + \dfrac{1}{2}X_2$, which has mean $\dfrac{1}{2}\mu_{X1} + \dfrac{1}{2}\mu_{X2}$, and since each X observation comes from the same distribution then $\mu_{X1} = \mu_{X2}$ so the mean is μ_X (it's unbiased). You can work it out when there are n observations.

Now the standard error of the mean is

$$\sqrt{\left(\frac{1}{2}\right)^2 \sigma_{X1}^2 + \left(\frac{1}{2}\right)^2 \sigma_{X2}^2 + 2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\sigma_{XY}} = \sqrt{\frac{1}{4}\sigma_{X1}^2 + \frac{1}{4}\sigma_{X2}^2} = \frac{1}{2}\sqrt{\sigma_{X1}^2 + \sigma_{X2}^2}.$$ The covariance can be set to zero

because we assume that we're making an independent random sample. Again since they come from the same distribution, $\sigma_{X1} = \sigma_{X2}$, the standard error is $\dfrac{1}{2}\sqrt{\sigma_X^2 + \sigma_X^2} = \dfrac{1}{2}\sqrt{2\sigma_X^2} = \dfrac{\sqrt{2}}{2}\sqrt{\sigma_X^2} = \dfrac{\sqrt{2}}{2}\sigma_X = \dfrac{1}{\sqrt{2}}\sigma_X$.

With n observations, the mean works out the same and the standard error of the average is

$$\sqrt{\left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}\sigma_x^2} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sigma_x^2} = \sqrt{\frac{n}{n^2}\sigma_x^2} = \frac{\sigma_X}{\sqrt{n}}.$$

Hypothesis Testing

Learning Outcomes (from CFA exam Study Session 3, Quantitative Methods)

Students will be able to:

- construct and interpret a confidence interval for a normally distributed random variable, and determine the probability that a normally distributed random variable lies inside a given confidence interval;

- define the standard normal distribution, explain how to standardize a random variable, and calculate and interpret probabilities using the standard normal distribution;

- explain the construction of confidence intervals;

- define a hypothesis, describe the steps of hypothesis testing, interpret and discuss the choice of the null hypothesis and alternative hypothesis, and distinguish between one-tailed and two-tailed tests of hypotheses;

- define and interpret a test statistic, a Type I and a Type II error, and a significance level, and explain how significance levels are used in hypothesis testing;

Hypothesis Testing

One of the principal tasks facing the statistician is to perform hypothesis tests. These are a formalization of the most basic questions that people ask and analyze every day – just contorted into odd shapes. But as long as you remember the basic common sense underneath them, you can look up the precise details of the formalization that lays on top.

The basic question 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.

A formal hypothesis sets a mathematical condition that I want to test. Often this condition takes the form of some parameter being zero for no relationship or no difference.

Statisticians tend to stand on their heads and ask: What if there were actually **no** relationship? (Usually they ask questions of the form, "suppose the conventional wisdom were true?") This statement, about "no relationship," is called the **Null Hypothesis**, sometimes abbreviated as H_o. 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?"

The "likelihood that I'm being fooled" is the p-value.

For a scientific experiment we typically first choose the level of certainty that we desire. This is called the significance level. This answers, "How low does the p-value have to be, for me to accept the formal hypothesis?" To be fair, it is important that we set this value first because otherwise we might be biased in favor of an outcome that we want to see. By convention, economists typically use 10%, 5%, and 1%; 5% is the most common.

A five percent level of a test is conservative, it means that we want to see so much evidence that there is only a 5% chance that we could be fooled into thinking that there's something there, when nothing is actually there. Five percent is not perfect, though – it still means that of every 20 tests where I decide that there is a relationship there, it is likely that I'm being fooled in one of those – I'm seeing a relationship where there's nothing there.

To help ourselves to remember that we can never be truly certain of our judgment of a test, we have a peculiar language that we use for hypothesis testing. If the "likelihood that I'm being fooled" is less than 5% then we say that the data allow us to *reject* the null hypothesis. If the "likelihood that I'm being fooled" is more than 5% then the data *do not reject* the null hypothesis.

Note the formalism: we never "accept" the null hypothesis. Why not? Suppose I were doing something like measuring a piece of machinery, which is supposed to be a centimeter long. The null hypothesis is that it is not defective and so is one centimeter in length. If I measure with a ruler I might not find any difference to the eye. So I cannot reject the hypothesis that it is one centimeter. But if I looked with a microscope I might find that it is not quite one centimeter! The fact that, with my eye, I don't see any difference, does not imply that a better measurement could not find any difference. So I cannot say that it is truly exactly one centimeter; only that I can't tell that it isn't.

Or again with the example of dice – the 6 might come up slightly more than 1/6 of the time, maybe if I rolled a million times I might finally distinguish a difference. But our hypothesis testing is much more limited, all we can say is that given the available tests we can't find a difference.

So too with statistics. If I'm looking to see if some portfolio strategy produces higher returns, then with one month of data I might not see any difference. So I would not reject the null hypothesis (that the new strategy is no improvement). But it is possible that the new strategy, if carried out for 100 months or 1000 months or more might show some tiny difference.

Not rejecting the null is saying that I'm not sure that I'm not being fooled. (Read that sentence again; it's not immediately clear but it's trying to make a subtle and important point.)

To summarize, Hypothesis Testing asks, "What is the chance that I would see the value that I've actually got, if there truly were no relationship?" If this p-value is lower than 5% then I reject the null hypothesis of "no relationship." If the p-value is greater than 5% then I do not reject the null hypothesis of "no relationship."

The rest is mechanics.

The null hypothesis would tell that a parameter has some particular value, say zero: $H_0 : \mu = 0$; the alternative hypothesis is $H_A : \mu \neq 0$. Under the null hypothesis the parameter has some distribution (often normal), so $H_0 : \mu \sim N\left(0, \sigma_{std\ err}\right)$. Generally we have an estimate for $\sigma_{std\ err}$, which is se (for small samples this inserts additional uncertainty). So I know that, under the null hypothesis, $\frac{\mu}{se}$ has a standard normal distribution (mean of zero and standard deviation of one). I know exactly what this distribution looks like, it's the usual bell-shaped curve:
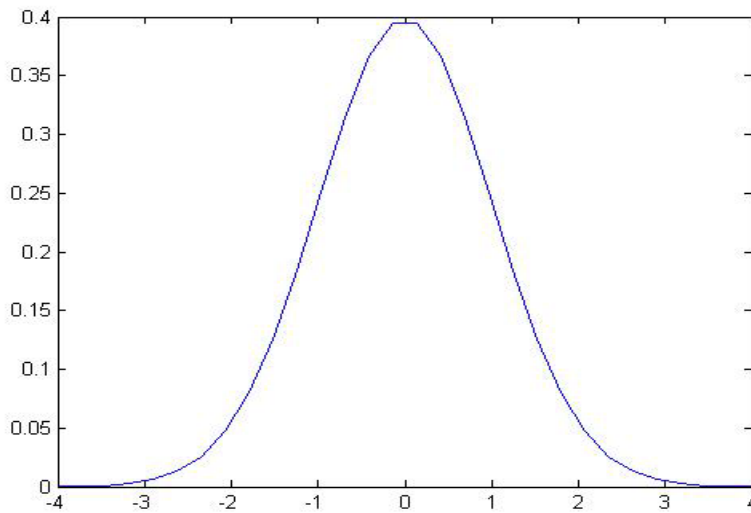
So from this I can calculate, "What is the chance that I would see the value that I've actually got, if there truly were no relationship?," by asking what is the area under the curve that is farther away from zero than the value that the data give. (I still don't know what value the data will give! I can do all of this calculation beforehand.)

A particular estimate of μ is generally going to be $\bar{X}$. So the test statistic is formed with $\dfrac{\bar{X}}{se}$.

Looking at the standard normal pdf, a value of the test statistic of 1.5 would not meet the 5% criterion (go back and calculate areas under the curve). A value of 2 would meet the 5% criterion, allowing us to reject the null hypothesis. For a 5% significance level, the standard normal **critical value** is 1.96: if the test statistic is larger than 1.96 (in absolute value) then its p-value is less than 5%, and vice versa. (You can find critical values by looking them up in a table or using the computer.)

Sidebar: Sometimes you see people do a one-sided test, which is within the letter of the law but not necessarily the spirit of the law (particularly in regression formats). It allows for less restrictive testing, as long as we believe that we know that there is only one possible direction of deviation (so, for example, if the sample could be larger than zero but never smaller). But in this case maybe the normal distribution is inapplicable. Personally whenever I read a paper where the authors do a one-sided test, I immediately become suspicious.

The test statistic can be transformed into measurements of μ or into a confidence interval.

If I know that I will reject the null hypothesis of $\mu = 0$ at a 5% level if the test statistic, $\dfrac{\bar{X}}{se}$, is greater than 1.96 (in absolute value), then I can change around this statement to be about $\bar{X}$. This says that if the estimated value of $\bar{X}$ is less than 1.96 standard errors from zero, we cannot reject the null hypothesis. So cannot reject if:

$$\frac{|\bar{X}|}{se} < 1.96$$

$$|\bar{X}| < 1.96 se$$

$-1.96se < \bar{X} < 1.96se$.

This range, $(-1.96se, 1.96se)$, is directly comparable to $\bar{X}$. If I divide $\bar{X}$ by its standard error then this ratio has a normal distribution with mean zero and standard deviation of one. If I don't divide then $\bar{X}$ has a normal distribution with mean zero and standard deviation, se .

If the null hypothesis is not zero but some other number, μ_{null} , then under the null hypothesis the estimator would have a normal distribution with mean of μ_{null} and standard error, se . To transform this to a standard normal would mean subtracting the mean and dividing by se , so cannot reject if $\dfrac{|\bar{X} - \mu_{null}|}{se} < 1.96$, i.e. cannot reject if $\bar{X}$ is within the range, $(\mu_{null} - 1.96se, \mu_{null} + 1.96se)$.

Confidence Intervals

We can use the same critical values to construct a confidence interval for the estimator, usually expressed in the form $\bar{X} \pm 1.96se$. This shows that, for a given sample size (therefore se , which depends on the sample size) that there is a 95% likelihood that the interval formed around a given estimator contains the true value.

This relates to hypothesis testing because if the confidence interval includes the null hypothesis then we cannot reject the null; if the null hypothesis value is outside of the confidence interval then we can reject the null.

Find p-values

We can also find p-values associated with a particular null hypothesis by turning around the process outlined above. If the null hypothesis is zero, then with a 5% significance level we reject the null if $\dfrac{\bar{X}}{se}$ is greater than 1.96 in absolute value. What if the ratio $\dfrac{\bar{X}}{se}$ were 2 – what is the smallest significance level that would still reject? (Check your understanding: is it more or less than 5%?)

We can compute the ratio $\dfrac{\bar{X}}{se}$ and then convert this number to a p-value, which is the smallest significance level that would still reject the null hypothesis (and if the null is rejected at a low level then it would automatically be rejected at any higher levels).

Type I and Type II Errors

Whenever we use statistics we must accept that there is a likelihood of errors. In fact we distinguish between two types of errors, called (unimaginatively) Type I and Type II. These errors arise because a null hypothesis could be either true or false and a particular value of a statistic could lead me to reject or not reject the null hypothesis, H_o. A table of the four outcomes is:

	H$_o$ is true	H$_o$ is false
Do not reject H$_o$	😀 good!	oops – Type II 💩
Reject H$_o$	oops – Type I 💩	😀 good!

Our chosen significance level (usually 5%) gives the probability of making an error of Type I. We cannot control the level of Type II error because we do not know just how far away H$_o$ is from being true. If our null hypothesis is that there is a zero relationship between two variables, when actually there is a tiny, weak relationship of 0.0001%, then we could be very likely to make a Type II error. If there is a huge, strong relationship then we'd be much less likely to make a Type II error.

There is a tradeoff (as with so much else in economics!). If I push down the likelihood of making a Type I error (using 1% significance not 5%) then I must be increasing the likelihood of making a Type II error.

Edward Gibbon notes that the emperor Valens would "satisfy his anxious suspicions by the promiscuous execution of the innocent and the guilty" (chapter 26). This rejects the null hypothesis of "innocence"; so a great deal of Type I error was acceptable to avoid Type II error.

Every email system fights spam with some sort of test: what is the likelihood that a given message is spam? If it's spam, put it in the "Junk" folder; else put it in the inbox. A Type I error represents putting good mail into the "Junk" folder; Type II puts junk into your inbox.

People play with setting the null hypothesis:

-        There is an advertisement for gas, "no other brand has been proven to be better";
-        Rand Paul offered a law that would allow a drug maker to publish any claim about drug efficacy that has not been proven false – does this mean that the claims will be true?;
-        Regulators of chemicals face this problem: policy of prohibit use of chemicals proved to be unsafe vs. policy of only allow chemicals proved to be safe.

Examples

Assume that the calculated average is 3, the sample standard deviation is 15, and there are 100 observations. The null hypothesis is that the average is zero. The standard error of the average is $se = \dfrac{15}{\sqrt{100}} = 1.5$. We can immediately see that the sample average is more than two standard errors away from zero so we can reject at a 95% confidence level.

Doing this step-by-step, the average over its standard error is $\dfrac{\overline{X}}{se} = \dfrac{3}{1.5} = 2$. Compare this to 1.96 and see that 2 > 1.96 so we can reject. Alternately we could calculate the interval, $(-1.96s, 1.96s)$, which is $\big((-1.96 \cdot 1.5), (1.96 \cdot 1.5)\big) = $ (-2.94, 2.94), outside of which we reject the null. And 3 is outside that interval. Or calculate a 95% confidence interval of $3 \pm 2.94 = (0.06, 5.94)$, which does not contain zero so we can reject the null. The critical value for the estimate of 3 is 4.55% (found from Excel either 2*(1-NORMSDIST(2)) if

using the standard normal distribution or 2*(1-NORMDIST(3,0,1.5,TRUE)) if using the general normal distribution with a mean of zero and standard error of 1.5).

If the sample average were -3, with the same sample standard deviation and same 100 observations, then the conclusions would be exactly the same.

Or suppose you find that the average difference between two samples, X and Y, (i.e. $\bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)$) is -0.0378. The sample standard deviation is 0.357. The number of observations is 652. These three pieces of information are enough to find confidence intervals, do t-tests, and find p-values.

How?

First find the standard error of the average difference. This standard error is 0.357 divided by the square root of the number of observations, so $\frac{.357}{\sqrt{652}} = 0.01398$.

So we know (from the Central Limit Theorem) that the average has a normal distribution. Our best estimate of its true mean is the sample average, -0.0378. Our best estimate of its true standard error is the sample standard error, 0.01398. So we have a normal distribution with mean -0.0378 and standard error 0.01398.

We can make this into a standard normal distribution by adding 0.0378 and dividing by the sample standard error, so now the mean is zero and the standard error is one.

We want to see how likely it would be, if the true mean were actually zero, that we would see a value as extreme as -0.0378. (Remember: we're thinking like statisticians!)

The value of -0.0378 is $\frac{-0.0378}{0.01398}$ = -2.70 standard deviations from zero.

From this we can either compare this against critical t-values or use it to get a p-value.

To find the p-value, we can use Excel just like in the homework assignment. If we have a standard normal distribution, what is the probability of finding a value as far from zero as -2.27, if the true mean were zero? This is 2*(1-NORMSDIST(-2.27)) = 0.6%. The p-value is 0.006 or 0.6%. If we are using a 5% level of significance then since 0.6% is less than 5%, we reject the null hypothesis of a zero mean. If we are using a 1% level of significance then we can reject the null hypothesis of a zero mean since 0.6% is less than 1%.

Or instead of standardizing we could have used Excel's other function to find the probability in the left tail, the area less than -0.0378, for a distribution with mean zero and standard error 0.01398, so 2*NORMDIST(-0.0378,0,0.01398,TRUE) = 0.6%.

Standardizing means (in this case) zooming in, moving from finding the area in the tail of a very small pdf, like this:
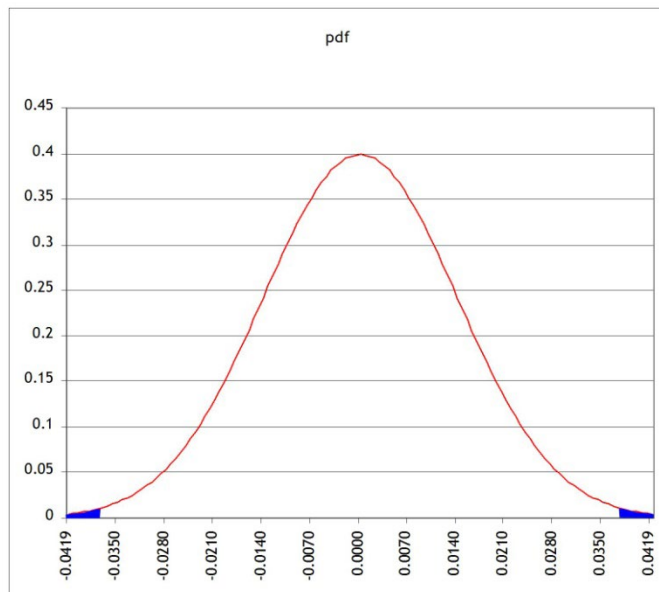
to moving to a standard normal, like this:
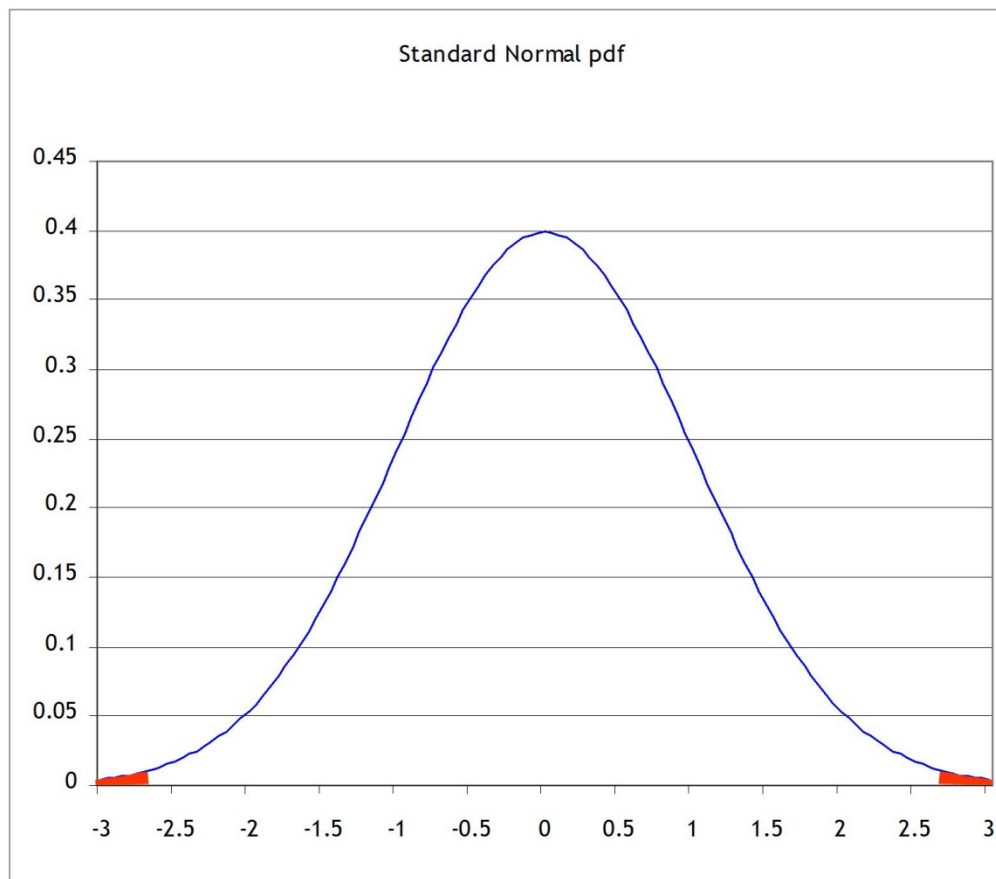
But since we're only changing the units on the x-axis, the two areas of probability are the same.

We could also work backwards. We know that if we find a standardized value greater (in absolute value) than 1.96, we would reject the null hypothesis of zero at the 5% level. (You can go back to your notes and/or HW1 to remind yourself of why 1.96 is so special.)

We found that for this particular case, each standard deviation is of size $\dfrac{.357}{\sqrt{652}} = 0.01398$. So we can multiply 1.96 times this value to see that if we get a value for the mean, which is farther from zero than 0.01398*1.96 = 0.0274, then we would reject the null. Sure enough, our value of -0.0378 is farther from zero, so we reject.

Alternately, use this 1.96 times the standard error to find a confidence interval. Choose 1.96 for a 95% confidence interval, so the confidence interval around -0.0378 is plus or minus 0.0274, -0.0378 ± 0.0274, which is the interval (-0.0652, -0.0104). Since this interval does not include zero we can be 95% confident that we can reject the null hypothesis of zero.

Sometimes we want to compare groups and ask, are they statistically significantly different from each other? Our formula that we learned previously has only one **n** – what do we do if we have two samples?

We want to figure out how to use the two separate standard errors to estimate the joint standard error; otherwise we'll use the same basic strategy to get our estimate, subtract off the null hypothesis (usually zero), and divide by its standard error. We just need to know, what is that new standard error?

To do this we use the sum of the two sample variances: if we are testing group 1 vs group 2, then a test of just group 1 would estimate its variance as $\dfrac{s_1^2}{n_1}$, a test of group 2 would use $\dfrac{s_2^2}{n_2}$, and a test of the group would estimate the standard error as $\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$.

We can make either of two assumptions: either the standard deviations are the same (even though we don't know them) or they're different (even though we don't know how different). It is more conservative to assume that they're different (i.e. don't assume that they're the same) – this makes the test less likely to reject the null.

Assuming that the standard errors are different, we compare this test statistic against a t-distribution with degrees of freedom of the minimum of either $(n_1 - 1)$ or $(n_2 - 1)$.

P-values

If confidence intervals weren't confusing enough, we can also construct equivalent hypothesis tests with p-values. Where the Z-statistic tells how many standard errors away from zero is the observed difference (leaving it for us to know that more than 1.96 implies less than 5%), the p-value calculates this directly. So a p-value for the difference above, between time spent by those with a college degree and those with an advanced degree, is found from -4.7919/1.6403 = -2.92. So the area in the tail to the left of -2.92 is NORMSDIST(-2.92) = .0017; the area in both tails symmetrically is .0034. The p-value for this difference is 0.34%; there is only a 0.34% chance that, if the true difference were zero, we could observe a number as big as -4.7919 in a sample of this size.

Confidence Intervals for Polls

I promised that I would explain to you how pollsters figure out the "±2 percentage points" margin of error for a political poll. Now that you know about Confidence Intervals you should be able to figure these

out. Remember (or go back and look up) that for a binomial distribution the standard error is $\sqrt{\dfrac{p(1-p)}{N}}$,

where p is the proportion of "one" values and N is the number of respondents to the poll. We can use our estimate of the proportion for p or, to be more conservative, use the maximum value of p(1 – p) where is p=

½. A bit of quick math shows that with $p = \dfrac{1}{2}$, $\sqrt{p(1-p)} = \sqrt{\dfrac{1}{2} \cdot \dfrac{1}{2}} = \dfrac{1}{2} = 0.5$. So a poll of 100 people has a

maximum standard error of $\dfrac{.5}{\sqrt{100}} = \dfrac{.5}{10} = .05$; a poll of 400 people has maximum standard error half that size, of .025; 900 people give a maximum standard error 0.0167, etc.

A 95% Confidence Interval is 1.96 (from the Standard Normal distribution) times the standard error; how many people do we need to poll, to get a margin of error of ±2 percentage points? We want

$1.96\sqrt{\dfrac{p(1-p)}{N}} < .02$ so this is, at maximum where p= ½, 2401.

A polling organization therefore prices its polls depending on the client's desired accuracy: to get ±2 percentage points requires between 2000 and 2500 respondents; if the client is satisfied with just ±5 percentage points then the poll is cheaper. (You can, and for practice should, calculate how many respondents are needed in order to get a margin of error of 2, 3, 4, and 5 percentage points. For extra, figure that a pollster needs to only get the margin to ±2.49 percentage points in order to round to ±2, so they can get away with slightly fewer.)

Here's a devious problem:

1. You are in charge of polling for a political campaign. You have commissioned a poll of 300 likely voters. Since voters are divided into three distinct geographical groups (A, B and C), the poll is subdivided into three groups with 100 people each. The poll results are as follows:

	total	A	B	C
number in favor of candidate	170	58	57	55
number total	300	100	100	100

Note that the standard deviation of the sample (not the standard error of the average) is given.

a. Calculate a t-statistic, p-value, and a confidence interval for the main poll (with all of the people) and for each of the sub-groups.

b. In simple language (less than 150 words), explain what the poll means and how much confidence the campaign can put in the numbers.

c. Again in simple language (less than 150 words), answer the opposing candidate's complaint, "The biased media confidently says that I'll lose even though they admit that they can't be sure about any of the subgroups! That's neither fair nor accurate!"

Complications from a Series of Hypothesis Tests

Often a modeler will make a series of hypothesis tests to attempt to understand the inter-relations of a dataset. However while this is often done, it is not usually done correctly. Recall from our discussion of Type I and Type II errors that we are always at risk of making incorrect inferences about the world based on our limited data. If a test has an significance level of 5% then we will not reject a null hypothesis until there is just a 5% probability that we could be fooled into seeing a relationship where there is none. This is low but

still is a 1-in-20 chance. If I do 20 hypothesis tests to find 20 variables that significantly impact some variable of interest, then it is likely that one of those variables is fooling me (I don't know which one, though). It is also likely that my high standard of proof meant that there are other variables which are more important but which didn't seem it.

Sometimes you see very stupid people who collect a large number of possible explanatory variables, run hundreds of regressions, and find the ones that give the "best-looking" test statistics – the ones that look good but are actually entirely fictitious. Many statistical programs have procedures that will help do this; help the user be as stupid as he wants.

Why is this stupid? It completely destroys the logical basis for the hypothesis tests and makes it impossible to determine whether or not the data are fooling me. In many cases this actually guarantees that, given a sufficiently rich collection of possible explanatory variables, I can run a regression and show that some variables have "good" test statistics – even though they are completely unconnected. Basically this is the infamous situation where a million monkeys randomly typing would eventually write Shakespeare's plays. A million earnest analysts, running random regressions, will eventually find a regression that looks great – where all of the proposed explanatory variables have test statistics that look great. But that's just due to persistence; it doesn't reflect anything about the larger world.

In finance, which throws out gigabytes of data, this phenomenon is common. For instance there used to be a relationship between which team won the Super Bowl (in January) and whether the stock market would have a good year. It seemed to be a solid result with decades of supporting evidence – but it was completely stupid and everybody knew it. Analysts still work to get slightly-less-implausible but still completely stupid results, which they use to sell their securities.

Consider the logical chain of making a number of hypothesis tests in order to find one supposedly-best model. When I make the first test, I have 5% chance of making a Type I error. Given the results of this test, I make the second test, again with a 5% chance of making a Type I error. The probability of not making an error on either test is (.95)(.95) = .9025 so the significance level of the overall test procedure is not 5% but 1 - .9025 = 9.75%. If I make three successive hypothesis tests, the probability of not making an error is .8574 so the significance level is 14.26%. If I make 10 successive tests then the significance level is over 40%! This means that there is a 40% chance that the tester is being fooled, that there is not actually the relationship there that is hypothesized – and worse, the stupid tester believes that the significance level is just 5%.

Issues with Canned Tests

Students often use a pre-written statistical test to declare that some difference is or is not statistically significant. This is a great efficiency! But it shouldn't come at the expense of understanding. What is being measured? To state that something is statistically significant is to state that it is "big" – so you'd better make sure that you know what in fact is big!

Let me give an example from an old exam. Take a moment to do this problem. In a medical study (reference below), people were randomly assigned to use either antibacterial products or regular soap. In total 592 people used antibacterial soap; 586 used regular soap. It was found that 33.1% of people using antibacterial products got a cold; 32.3% of people using regular soap got colds.
   a.   Test the null hypothesis that there is no difference in the rates of sickness for people using regular or antibacterial soap. (What is the p-value?)
         Standard deviation : sqrt(p*(1-p)) : sqrt(.331*(1-.331))
         Standard error: sqrt(p(1-p))/sqrt(n) : sqrt(.331(1-.331))/sqrt(592)
         Difference .331 - .323

$$\text{Standard error of difference: } \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

b.   Create a 95% confidence interval for the difference in sickness rates. What is the 90% confidence interval? The 99% interval?

E.L.Larson, S.X. Lin, C. Gomez-Pichardo, P. Della-Latta, (2004). "Effect of Antibacterial Home Cleaning and Handwashing Products on Infectious Disease Symptoms: A Randomized Double-Blind Trial," Ann Intern Med, 140(5), 321-329.

Many students obliged by forming a statistical test to show whether there was a significant difference, but without ever noticing the counter-intuitive direction! In this case a test of statistical significance is useless and irrelevant – certainly you don't need to do any calculations to assert that this study shows no beneficial effect of antibacterial soap!

On many homework assignments, I've observed similar answers. Students rush into the mechanics of the test without any assessment. A statistical test is an important component of an argument but it is not the alpha and omega. Much more of the time and mental effort needs to go into thinking about the other factors – why might you observe these values? Have you got the right measure in the first place? Have you got a reasonable sample? What are some of the possible hypotheses that explain the difference? Is there a way to eliminate some of these hypotheses or to reduce the variation?

Once you've done the hard thinking and got an interesting measure, you can ask whether it is statistically significant. And this class will help you be more adroit with those tests.

Bayesian Stats

A reminder about basic stats – and illustration of the power of Bayesian statistics.

We did this example before: a 99% accurate test reveals that a person tests positive for a disease. How likely does the patient actually have the disease?

It depends.

If population overall has prevalence of 0.1%, then testing 1000 people will find the one person with disease plus 10 who don't have it (1% error of 99% test; 1% of 1000 = 10) – so a positive test for the disease means a 1/11 chance of actually having it.

On the other hand, if a subgroup of the population has a higher prevalence (say 1%) then putting together this prior information with the fact of a positive test implies that a positive test means about a 50% chance that the patient actually has the disease (10 people who have it plus 10 false positives).

So in the first case, the expected value of whether the person has the disease is 0.09 (=1/11); in the second case the expected value is 0.5. So the expected value depends on the empirical information (positive test result) but also the prior expectation (what is your guess of prevalence in subgroup).

In much of stats you can see this tradeoff between data and prior. In this case, with one data point, the prior is very important. With more data the importance of the prior recedes, but there are many important cases where people's priors remain a key determinant.

Details of Distributions T-distributions, chi-squared, etc.

Take the basic methodology of Hypothesis Testing and figure out how to deal with a few complications.

T-tests

The first complication is if we have a small sample and we're estimating the standard deviation. In previous examples, we used a large sample. For a small sample, the estimation of the standard error introduces some additional noise – we're forming a hypothesis test based on an estimation of the mean, using an estimation of the standard error.

How "big" should a "big" sample be? Evidently if we can easily get more data then we should use it, but there are many cases where we need to make a decision based on limited information – there just might not be that many observations. Generally after about 30 observations is enough to justify the normal distribution. With fewer observations we use a t-distribution.

To work with t-distributions we need the concept of "Degrees of Freedom" (df). This just takes account of the fact that, to estimate the sample standard deviation, we need to first estimate the sample average, since the standard deviation uses $\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$. So we don't have as many "free" observations. You might remember from algebra that to solve for 2 variables you need at least two equations, three equations for three variables, etc. If we have 5 observations then we can only estimate at most five unknown variables such as the mean and standard deviation. And "degrees of freedom" counts these down.

If we have thousands of observations then we don't really need to worry. But when we have small samples and we're estimating a relatively large number of parameters, we count degrees of freedom.

The family of t-distributions with mean of zero looks basically like a Standard Normal distribution with a familiar bell shape, but with slightly fatter tails. There is a family of t-distributions with exact shape depending on the degrees of freedom; lower degrees of freedom correspond with fatter tails (more variation; more probability of seeing larger differences from zero).

This chart compares the Standard Normal PDF with the t-distributions with different degrees of freedom.
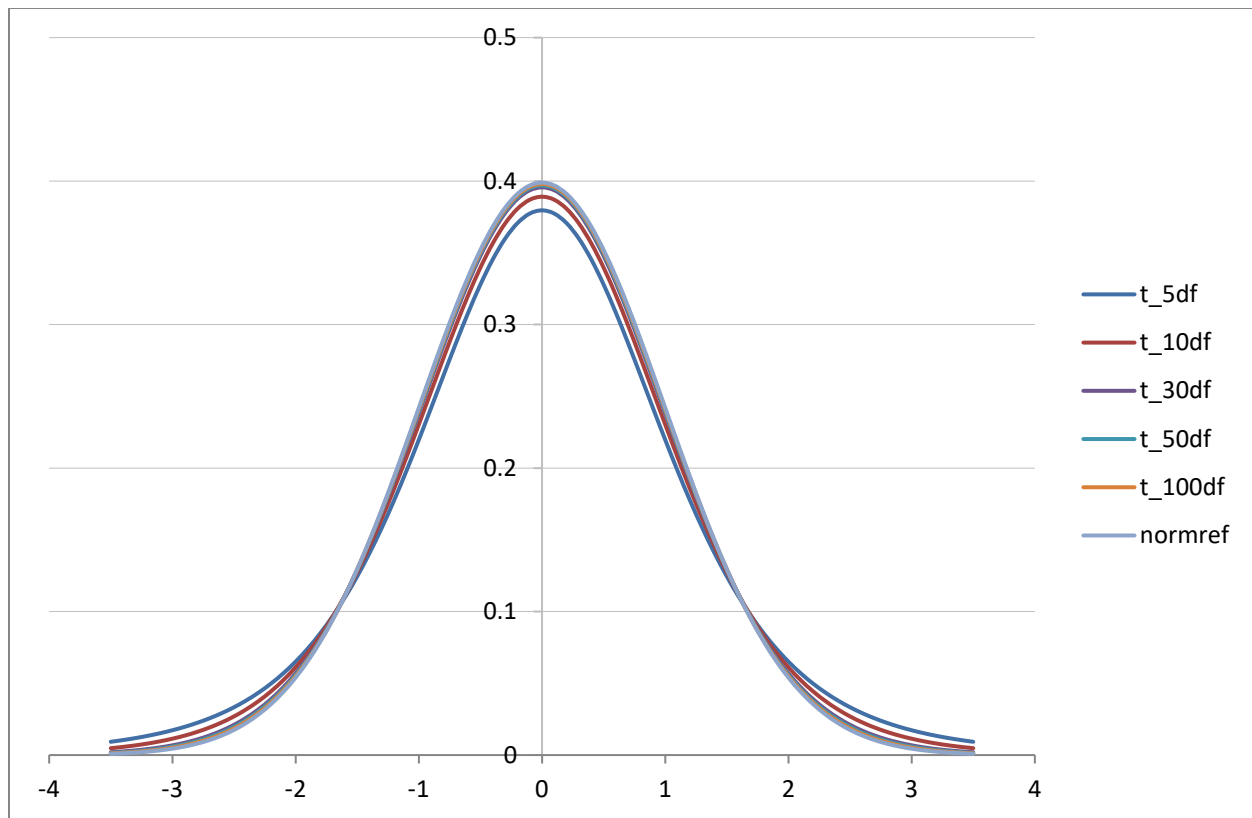
This table shows the different critical values to use in place of our good old friend 1.96:

Critical Values for t vs N

df	95%	90%	99%
5	2.57	2.02	4.03
10	2.23	1.81	3.17
20	2.09	1.72	2.85
30	2.04	1.70	2.75
50	2.01	1.68	2.68
100	1.98	1.66	2.63
Normal	1.96	1.64	2.58

The higher numbers for lower degrees of freedom mean that the confidence interval must be wider – which should make intuitive sense. With just 5 or 10 observations a 95% confidence interval should be wider than with 1000 or 10,000 observations (even beyond the familiar sqrt(N) term in the standard error of the average).

T-tests with two samples

When we're comparing two sample averages we can make either of two assumptions: either the standard deviations are the same (even though we don't know them) or they could be different. It is more conservative to assume that they're different (i.e. don't assume that they're the same) – this makes the test less likely to reject the null.

Assuming that the standard errors are different, we compare this test statistic against a t-distribution with degrees of freedom of the minimum of either $(n_1 - 1)$ or $(n_2 - 1)$.

Sometimes we have paired data, which can give us more powerful tests.

We can test if the variances are in fact equal, but a series of hypothesis tests can give us questionable results.

Note on the t-distribution:

Talk about the t distribution always makes me thirsty. Why? It was originally called "Student's t distribution" because the author wanted to remain anonymous and referred to himself as just a student of statistics. William Gosset worked at Guinness Brewing, which had a policy against its employees publishing material based on their work – they didn't want their brewing secrets revealed! It's amusing to think that Gosset, who graduated top of his class from the one of the world's top universities in 1899, went to work at Guinness – although at the time that was a leading industrial company doing cutting-edge research. A half-century later, the brightest students from top universities would go to GM; after a century the preferred destinations would be Google or Goldman Sachs. The only thing those companies have in common is the initial G.

Other Distributions

There are other sampling distributions than the Normal Distribution and T-Distribution. There are χ^2 (Chi-Squared) Distributions (also characterized by the number of degrees of freedom); there are F-Distributions with two different degrees of freedom. For now we won't worry about these but just note that the basic procedure is the same: calculate a test statistic and compare it to a known distribution to figure out how likely it was, to see the actual value.

(On Car Talk they joked, "I once had to learn the entire Greek alphabet for a college class. I was taking a course in … Statistics!")

Simple Machine Learning

From basic notions of mean and standard deviation, we can quickly move to some simple types of machine learning. This is a great example of a very simple idea that has some fancy-sounding terminology. The idea is that if you want to classify a new observation then the easiest guess is to ask how old observations that were very near were classified. "Birds of a feather flock together," or "You're judged by the company you keep."

There are many possibilities, where we gather data on some preliminary values and try to predict something else. If we have a big dataset on past students who were admitted or not to a certain program, we could use this data to predict future admits. Lots of marketing might use this sort of algorithm: if consumers are similar by some characteristics then they might be similarly receptive to a certain ad.

With the PUMS data, we can try to predict which borough of the city a person lives in. Note that this is the inverse of the problem that is often posed, "oh I can tell you're from Brooklyn," because... This is looking at the opposite: are there certain characteristics that allow us to predict what borough a person lives in?

Alternately, the Consumer Expenditure Survey data has information on the type of heat used (coded as 1 natural gas, 2 electricity, 3 oil, 4 other, 5 none). We want to guess a new observation based on households that are "near".

The machine learning technique called "K Nearest Neighbors" or "k-nn" uses other observations that are "nearby" to try to classify new observations.

What does "near" mean? If we have a list of numeric data then the temptation is to just use simple distance (typically Euclidean). There are two aspects to this choice: first, what variables are helpful in classification; second, how are these variables scaled? The choice of what variables is a bit tricky since we want to find some good ones but not too many (too many, relative to the size of the data, gets into the "curse of dimensionality" and there are usually few neighbors). That usually requires a bit of background knowledge – this is called "machine learning" but it's actually strongly human-controlled machine learning (so cyborg learning?).

The second part of "near" is a bit more subtle: the scaling of each variable is important. If a college is classifying high school students as either admit or not, they might use GPA and SAT. If HS GPA is on a scale of 0-4 then for selective colleges most of the relevant admissions will have GPA from 3.5-4. SAT scores on the other hand (for now assume they're math plus verbal) could have differences of hundreds of points. So the SAT score variation will swamp the GPA variation. (This is why some people think of scores on standardized tests as their percentile.)

Detour on Ranking

We often see statistics reported that rank a number of different units based on a number of different measures. For instance, these could be the US News ranking of colleges, or magazine rankings of city livability, or sports rankings of college teams, or any of a multitude of different things. We would hope that statistics could provide some simple formulas; we would hope in vain.

Education: College rankings try to combine student/faculty ratios, measures of selectivity, SAT scores, GPA; some add in numbers of bars near campus or the prestige of journals in which faculty publish. What is best? School teachers face efforts to rank them, by student test score improvements as well as other factors; schools and districts are ranked by a variety of measures.

Sports might seem to have it relatively easy since there is a single ranking given by pre-arranged rules, but still fans can argue: a team has a good offense because they scored a lot (even though some other team won more games); some players are better on defense but worse on offense. Sports Illustrated tried to rank the 100 all-time best sports stars, somehow comparing baseball player Babe Ruth with the race horse Secretariat! Most magazines know that rankings drive sales and give buzz.

Food nutrition trades off calories, fat content, fiber, vitamin and mineral content; who is to say whether kale or blueberries are healthier? Aren't interaction effects important? Someone trying to lose weight would make a very different ranking than someone training for a marathon.

Sustainability or "green" rankings are difficult: there are so many trade-offs! If we care about global warming then we look at CO_2 emissions, but what about other pollutants? Is nuclear power better than natural gas? Ethical consumption might also consider the material conditions of workers (fair-trade coffee or no-sweatshop clothing) or other considerations.

Politics: which political party is better for the economy? Could measure stock returns or unemployment rate or GDP growth or hundreds of others. Average wage or median earnings (household or individual)? Each set of measures could give different results. You can try this yourself, get some data from FRED (http://research.stlouisfed.org/fred2/) and go wild.

In the simplest case, if there is just a single measured variable, we can rank units based on this single measure, however even in this case there is rarely a clear way of specifying which rankings are based on differences that are large and which are small. (The statistical theory is based on "order statistics.") If the

outcome measure has, for example, a normal distribution, then there will be a large number of units with outcomes right around the middle, so even small measurement errors can make a big difference to ranking.

In the more complicated (and more common) case, we have a variety of measures of outcomes and want to rank units based on some amalgamation of these outcomes. A case where a large number of inputs generates a single unit output looks like a utility function from micro theory: I face a choice of hundreds (or thousands) of different goods, which I put into a single ranking: I say that the utility of some bundle of goods is higher than the utility of some other bundle and so would rank it higher (even if both were affordable).

However there is no way to generate a composite utility function for a group of people that completely and successfully takes account of the information of individual choices! (This result is due to CCNY alumnus and Nobel Laureate Ken Arrow.)

In general many rankings can be substantially changed by adding factors or even changing the units of certain of the factors (changing the measure of "near" as discussed before).

Many rankings take an equal weighting of each item, but there is absolutely no good reason to do this generally: why would we believe that each measure is equally valid? Some rankings might arbitrarily choose weights or take a separate survey to find weights (equally problematic!). You could average what fraction of measures achieve some hurdle.

One possible way around this problem is to just ask for people's rankings (let them figure out what weights to use in their own utility functions) and report some aggregation. However here again there is no single method that is guaranteed to give correct aggregations (this is the Ken Arrow result again). Some surveys ask people to rank units from 1-20, then add the rankings and the unit with the lowest number wins. But what if some people rank number 1 as far ahead of all of their competitors, while others see the top 3 as tight together? This distance information is omitted from the rankings. Some surveys might, instead, give 10 points for a #1 ranking, 8 points for #2, and so on – but again this presupposes some distance between the ranks.

This is not to say that ranking is hopeless or never informative, just that there is no single path that will inerrantly give the correct result. Working through various rankings, an analyst might determine that a broad swathe of weights upon the various measures would all give similar rankings to certain outliers. It would be useful to know that a particular unit is almost always ranked near the top while some other one is nearly always at the bottom.

As economists we must also think about the game theory around these rankings: there will usually be a dynamic game underway. If a prominent publication ranks colleges by some set of numbers, then lower-ranked colleges will try to change their numbers to improve their rank. There are a variety of ways to do this, in a range from honest to nefarious (historically many simply lied, since there was essentially a zero penalty to dishonesty). High schools do this when evaluated based on test scores.

Cathy O'Neil's new book, *Weapons of Math Destruction*, gives many more examples of problems that arise.

Other Ignorant Beliefs

While I'm working to extirpate popular heresies, let me address another one, which is particularly common when the Olympics roll around: the extraordinary belief that outliers can give useful information about the average value. We hear these judgments all of the time: some country wins an unusual number of Olympic medals, thus the entire population of the country must be unusually skilled at this task. Or some gender/race/ethnicity is overrepresented in a certain profession thus that gender/race/ethnicity is more skilled on average. Or a school has a large number of winners of national competitions, thus the average is higher. Really?

Statistically speaking, the extreme values of a distribution depend on many parameters such as the higher moments. If I have two distributions with the exact same mean, standard deviation, and skewness, but different values of kurtosis, then one distribution will systematically have higher extremes (by definition of kurtosis). So in general it is not true to infer that a higher number of extreme values implies a higher mean. But people do.

Rankings can be shifted by different values of "near" as can machine learning algorithms. It is up to you to learn about how to use these most adroitly.

The variation in a measure is sometimes called its "information". Consider even a simple case where students' grades in a class are determined by even weighting of 2 exams. If scores on one exam are much more variable than scores on the other exam then they don't actually end up contributing equal weight to student ranking. (Think of the limiting case where everyone gets the same score on one exam, therefore it has no contribution to ranking even if it is given 50% weight.)

A common way to manage this is to standardize the predictors (subtract mean and divide by standard deviation) or scale them to be all in the [0,1] interval, although this is far from perfect. There is an art to choosing predictors. Although it might not seem obvious, this is essentially the same problem as with rankings.

Below is a simple program that you can modify and improve. It tries to classify what borough the people in NYC live in. I leave plenty of room for you to improve and begin by just using the predictors of a person's income and how much they pay for their housing (whether rented or owned).

It uses a technique that we'll often return to: splitting the data into a training set and a test set. If the point of a model is to predict some data, then I want to test it out on some data that was not used for training. For example you've doubtless taken classes with various types of exams. Sometimes the instructor will give students a number of practice problems then the exam would consist of some of those problems. Other times the instructor will give practice problems but then the exam is new problems that are related to the practice but not identical. I think you'd agree that the second type is more difficult!

We want to test our models similarly and don't just reuse data to give an easy test. We take out some of the data and don't use that in the estimation. The data used for estimation is the "training" data, that we use to train the model. The test data is separate, used to test how well that model performs on data that it hasn't seen before. Here we use 80% of the data as the training set and the remaining 20% as the test set.

The "set seed" command is a bit of magic that lets us take a random sample but if you do it again the computer would take the same "random" sample. The computer doesn't actually take a random sample but it is actually pseudo-random where complicated algorithms create numbers that look random in many ways but are actually deterministic so if we start from the same value then we get the same list of random numbers. The "seed" sets that starting point. You might think, why not just take the first 80% of the sample, but that would depend on the assumption that the ordering of data is random. Many datasets have structure so the observations might be ordered in some way.

The program finishes by using the "knn" routine from the "class" package (for various classification algorithms). It can use different numbers of nearest neighbors so experiments with using 1, 3, 5, 7 or 9 nearest neighbors for the classification and reports how accurate each one is.

```
norm_varb <- function(X_in) {
```

```
    (X_in - mean(X_in, na.rm = TRUE))/sd(X_in, na.rm = TRUE)
}

# classification problem: of people in NYC, can we classify which borough?
# subset by in_NYC then create a factor for borough; consider only working
age since use income
dat_NYC <- subset(acs2017_ny, (acs2017_ny$in_NYC == 1)&(acs2017_ny$AGE >
20)&(acs2017_ny$AGE < 66))
attach(dat_NYC)
borough_f <- factor((in_Bronx + 2*in_Manhattan + 3*in_StatenI +
4*in_Brooklyn + 5*in_Queens), levels=c(1,2,3,4,5),labels =
c("Bronx","Manhattan","Staten Island","Brooklyn","Queens"))

# what variables do we think are relevant in classifying by borough?
# >>NOT<< PUMA since neighborhood likely perfectly classifies...
# try income_total, owner_cost combined with rent_cost

housing_cost <- OWNCOST + RENT
norm_inc_tot <- norm_varb(INCTOT)
norm_housing_cost <- norm_varb(housing_cost)

data_use <- data.frame(norm_inc_tot,norm_housing_cost)
good_obs_data_use <- complete.cases(data_use,borough_f)
dat_use <- subset(data_use,good_obs_data_use)
y_use <- subset(borough_f,good_obs_data_use)
detach(dat_NYC)


set.seed(12345)
NN_obs <- sum(good_obs_data_use == 1)
select1 <- (runif(NN_obs) < 0.8)

train_data <- subset(dat_use,select1)
test_data <- subset(dat_use,(!select1))
cl_data <- y_use[select1]
true_data <- y_use[!select1]

summary(cl_data)
prop.table(summary(cl_data))
summary(train_data)

require(class)
for (indx in seq(1, 9, by= 2)) {
  pred_borough <- knn(train_data, test_data, cl_data, k = indx, l = 0, prob
= FALSE, use.all = TRUE)

  num_correct_labels <- sum(pred_borough == true_data)
  correct_rate <- num_correct_labels/length(true_data)
  print(c(indx,correct_rate))

}

# is this good?  If you just guessed randomly, how many would you get?
# how much better can you do?
```

(Let me crush a bit, I learned much of this from the great book *Doing Data Science* by Cathy O'Neil & Rachel Schutt – get it, read it, love it!)

Jumping into OLS

OLS is Ordinary Least Squares, which as the name implies is ordinary, typical, common – something that is widely used in just about every economic analysis.

We are accustomed to looking at graphs that show values of two variables and trying to discern patterns. Consider again 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 Jan 2, 1991 to Jan 31, 2006)
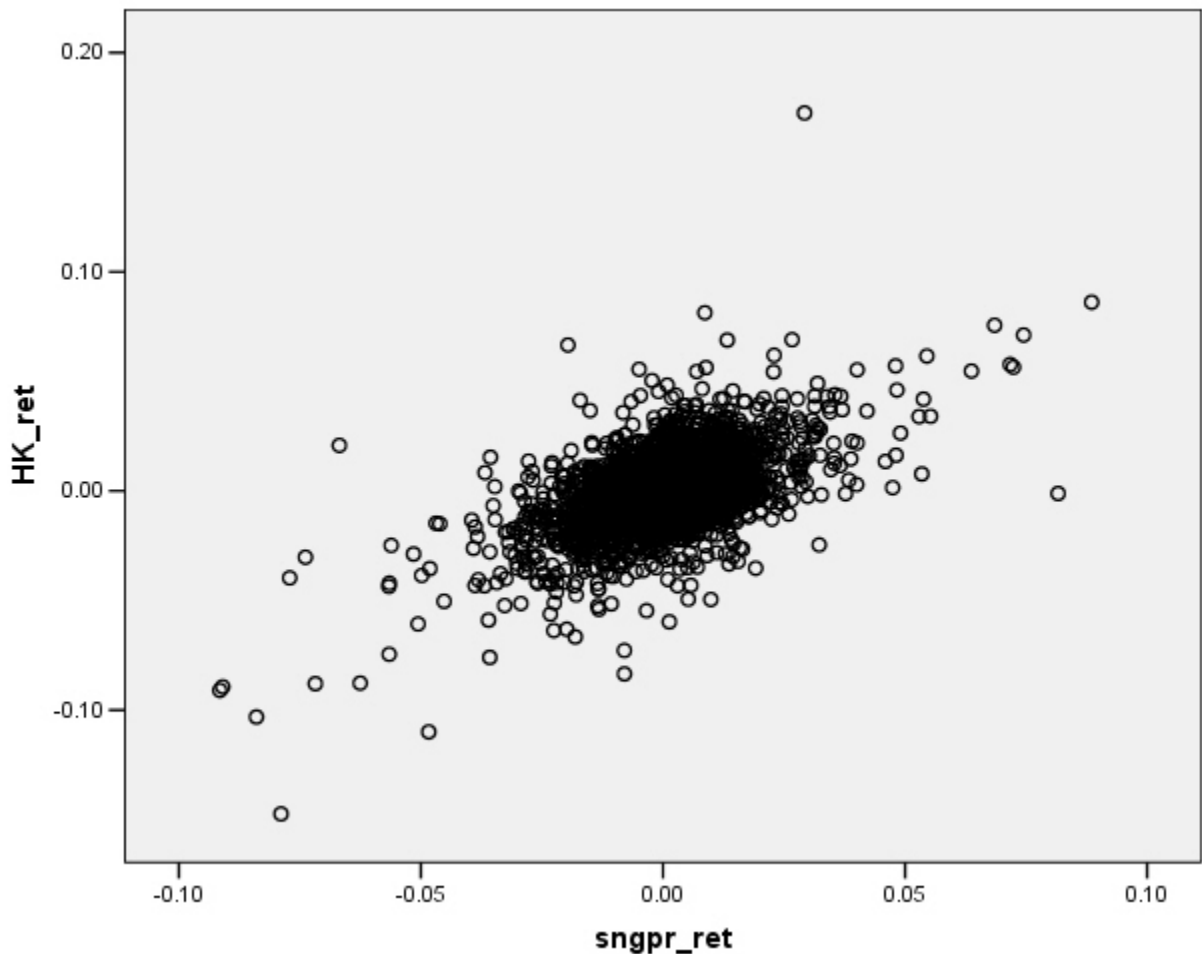
This next graph shows the S&P 500 returns and interest rates (1-month Eurodollar) during 1989-2004.
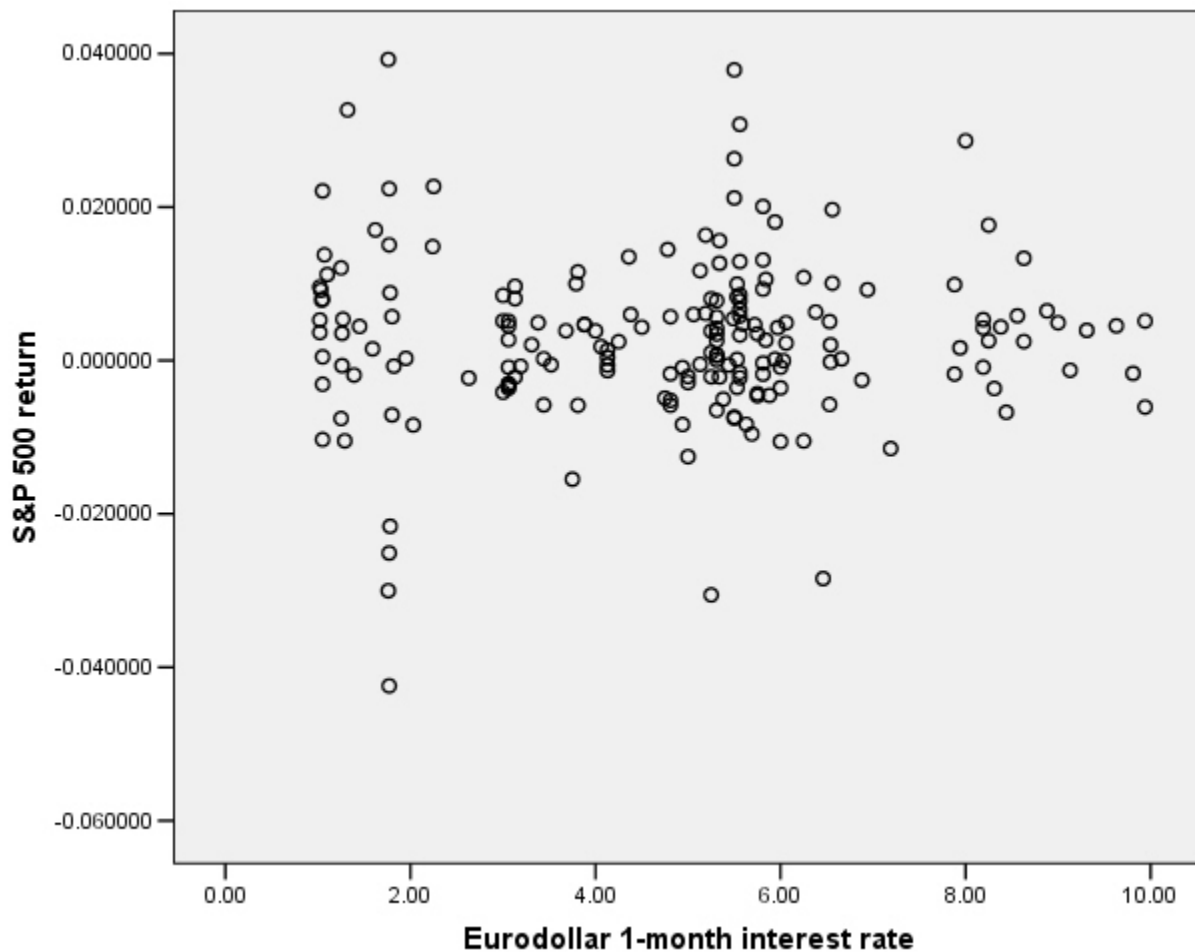
You don't have to be a highly-skilled econometrician to see the difference in the relationships. It would seem reasonable that the Hong Kong and Singapore stock indexes are closely linked while the US stock index is not closely related to interest rates.

So 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? How can we try to guard against seeing relationships where, in fact, none actually exist? We will consider each of these questions in turn.

How can we measure the relationship?

Facing a graph like the Hong Kong/Singapore stock indexes, we might represent the relationship by drawing a line, something like this:
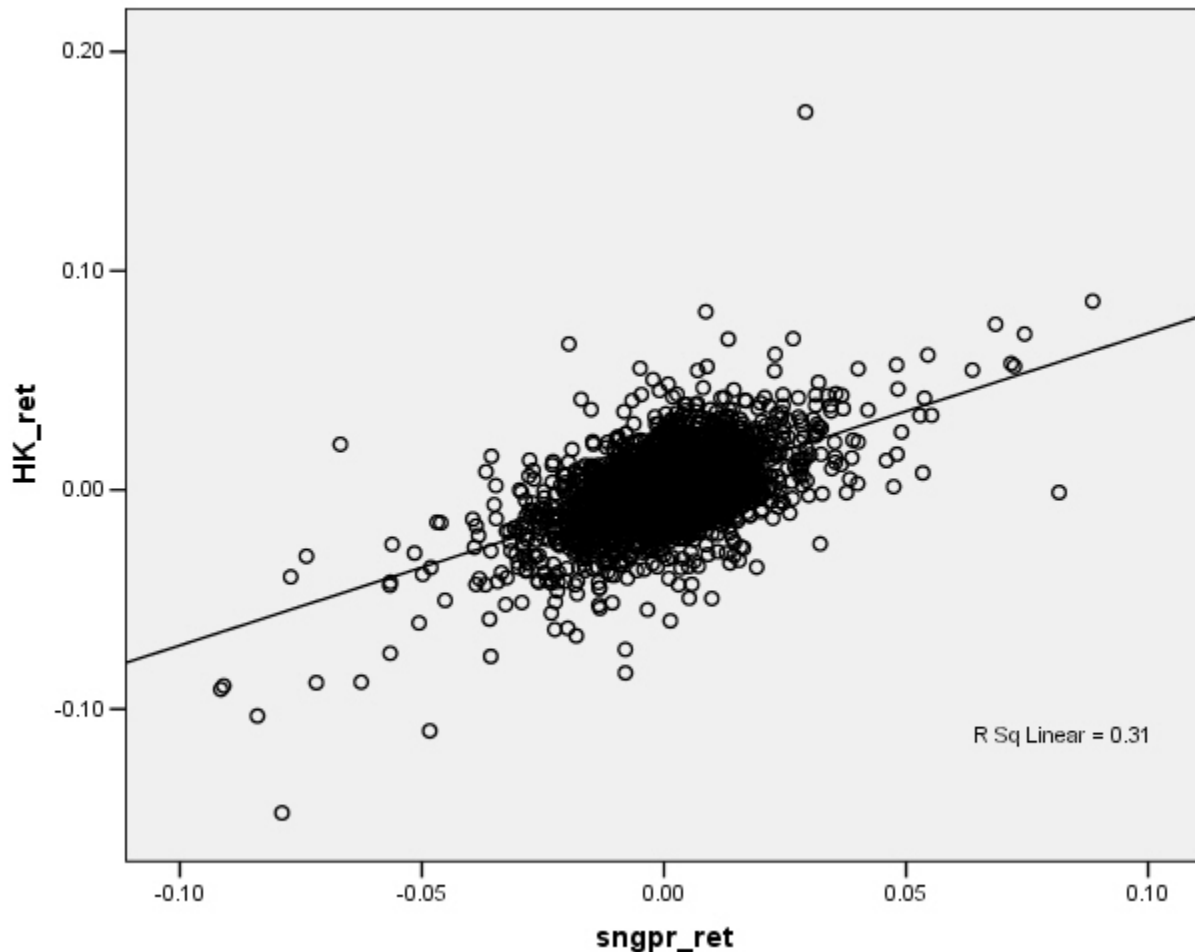
Now if this line-drawing were done just by hand, just sketching in a line, then different people would sketch different lines, which would be clearly unsatisfactory. What is the process by which we sketch the line?

Typically we want to find a relationship because we want to predict something, to find out that, if I know one variable, then how does this knowledge affect my prediction of some other variable. We call the first variable, the one known at the beginning, X. The variable that we're trying to predict is called Y. So in the example above, the Singapore stock index is X and the Hong Kong index is Y. The line that we would draw in the picture would represent our best guess of what Y would be, given our knowledge about X.

This line is drawn to get the best guess "close to" the actual Y values – where by "close to" we actually minimize the average squared distance. Why square the distance? This is one question which we will return to, again and again; for now the reason is that a squared distance really penalizes the big misses. If I square a small number, I get a bigger number. If I square a big number, I get a HUGE number. (And if I square a number less than one, I get a smaller number.) So minimizing the squared distance will mean that I am willing to make a bunch of small errors in order to reduce a really big error. This is why there is the "LS" in "OLS" -- "Ordinary Least Squares" finds the least squared difference.

A computer can easily calculate a line that minimizes the squared distance between each Y value and the best prediction. There are also formulas for it. (We'll come back to the formulas; put a lightning bolt here to remind us: ⚡ .)

For a moment consider how powerful this procedure is. A line that represents a relationship between X and Y can be entirely produced by knowing just two numbers: the y-intercept and the slope of the line. In algebra class you probably learned the equation as:

$$Y = mX + b$$

where the slope is m and the y-intercept is b. When $X = 0$ then $Y = b$, which is the value of the line when the line intersects the Y-axis (when X is zero). The y-intercept can be positive or negative or zero. The slope is the value of $\frac{\Delta Y}{\Delta X}$, which tells how much Y changes when X changes by one unit. To find the predicted value of Y at any point we substitute the value of X into the equation.

In econometrics we will typically use a different notation,

$$Y = \beta_0 + \beta_1 X$$

where now β_0 is the y-intercept and the slope is β_1. (Econometricians looooove Greek letters like beta, get used to it!)

The relationship between X and Y can be positive or negative. Basic economic theory says that we expect that the amount demanded of some item will be a positive function of income and a negative function of price (for a normal good). We can easily have a case where $\beta_1 < 0$.

If X and Y had no systematic relation, then this would imply that $\beta_1 = 0$ (in which case, β_0 is just the mean of Y). In the $\beta_1 = 0$ case, Y takes on higher or lower values independently of what is the level of X.

This is the case for the S&P 500 return and interest rates:

R Sq Linear = 2.018E-4

R Sq Linear = 2.018E-4

R Sq Linear = 2.018E-4
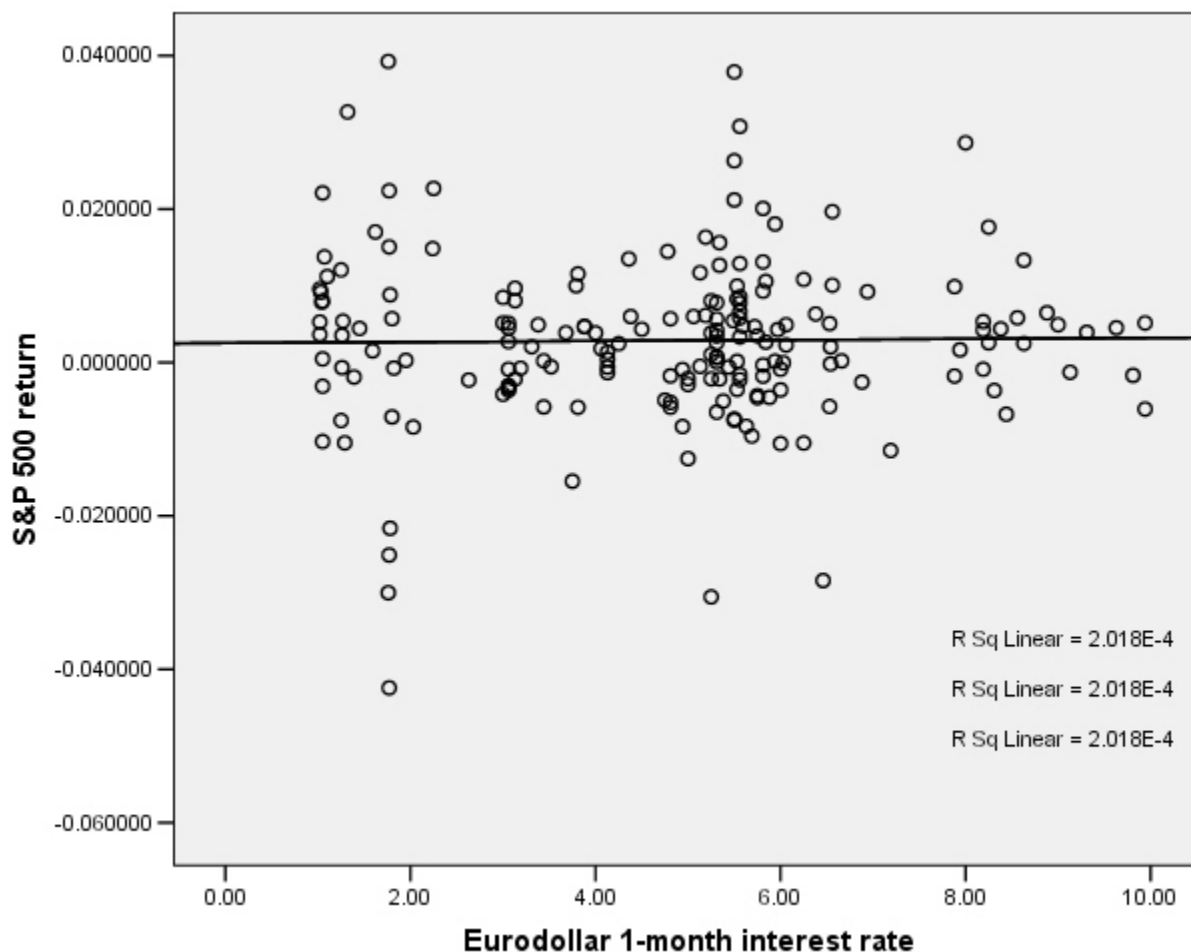
So there does not appear to be any relationship.

Let's fine up the notation from above a bit more: when we fit a line to the data, we do not always have Y exactly and precisely equal to $\beta_0 + \beta_1 X$. Sometime Y is a bit bigger, sometimes a bit smaller. The difference is an error in the model. So we should actually write $Y = \beta_0 + \beta_1 X + \varepsilon$ where epsilon is the error between the model value of Y and the actual observed value.

Computer programs will easily compute this OLS line; even Excel will do it. When you create an XY (Scatter) chart, then right-click on the data series, "Add Trendline" and choose "Linear" to get the OLS estimates.

Other Notation:

There is another possible notation, that $Y = \alpha + \beta X$. This is often implicit in discussions of hedge funds or financial investing. If X is the return on the broad market (the S&P500, for example) and Y is the return of a hedge fund, then the hedge fund managers must make a case that they can provide "alpha" – that for their hedge fund $\alpha > 0$. This implies that no matter what the market return is, the hedge fund will return better. The other desirable case is for a hedge fund with beta near zero – which might seem odd at first. But this provides diversification: a low beta means that the fund returns do not really depend on the broader market. An investment with a zero beta and alpha of 0.5% is a savings account. An investment promising zero beta and alpha of 20% is a fraud. Beta equal to 1 and alpha equal to -0.2% is an index fund.

This representation is powerful because it neatly and compactly summarizes a great deal of underlying variation. Consider the case of looking at the time that people spend eating and drinking, as reported in the ATUS data; we want to see if there is a relationship with the person's age. If we compute averages for each age (average time spent by people who are 18 years old, average time spent by people who are 19 years old, 20 years old, etc – all the way to 85 years old) along with the standard deviations we get this chart:

Age vs Minutes Eating

Age
from ATUS 2003-2013
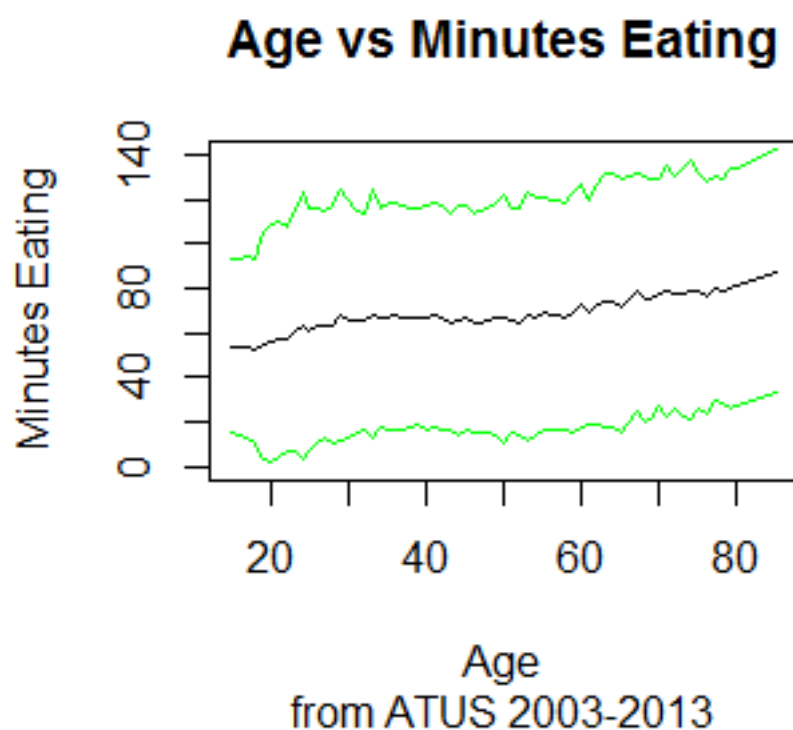
There seems to be an upward trend although we might distinguish a flattening of time spent, between ages 30 and 60. But all of this information takes a table of numbers with 67 rows and 4 columns so 268 separate numbers! If we represent this as just a line then we need just two numbers, the intercept and the slope. This also makes more effective use of the available information to "smooth out" the estimated relationship. (For instance, there is a leap up for 29-year-olds but then a leap back down – do we really believe that there is really that sort of discontinuity or do we think this could just be the randomness of the data? A fitted line would smooth out that bump.)

Angrist & Pischke distinguish the Conditional Expectation Function as the average value of Y given some X; and OLS is simply the best linear predictor.

How can we distinguish cases in the middle?

Hopefully you've followed along so far, but are currently wondering: How do I tell the difference between the Hong Kong/Singapore case and the S&P500/Interest Rate case? Maybe art historians or literary theorists can put up with having "beauty" as a determinant of excellence, but what is a beautiful line to econometricians?
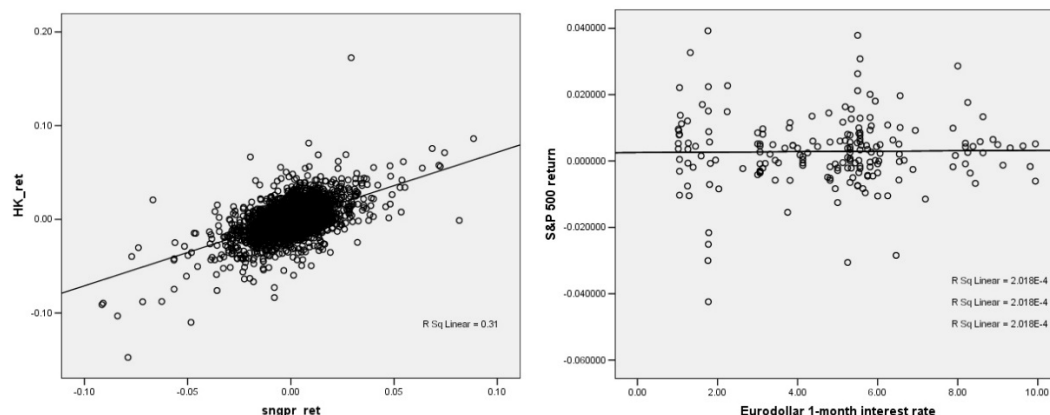
There are two separate answers here, and it's important that we separate them. Many analyses muddle them up. One answer is simply whether the line tells us useful information. Remember that we are trying to estimate a line in order to persuade (ourselves or someone else) that there is a useful relationship here. And "useful" depends crucially upon the context. Sometimes a variable will have a small but vital relationship; others may have a large but much less useful relation. To take an example from macroeconomics, we know that the single largest component of GDP is consumption, so consumption has a large impact on GDP. However US consumption is based on the individual choices of 300m people, so it's difficult for policymakers to have a direct and immediate effect upon it. Beginning students are often surprised to discover how important an effect inventory investment has historically had on US GDP growth, even though inventory adjustments are a tiny slice of GDP. The Fed's actions have a tiny direct effect yet we all agree that they are very important because this tiny effect may help the economy in huge ways.

This first question, does the line persuade, is always contingent upon the problem at hand; there is no easy answer. You can only learn this by reading other people's analyses and by practicing on your own. It is an art form to be learned, but the second part is science.

The economist Dierdre McCloskey has a simple phrase, "How big is big?" This is influenced by the purpose of the research and the aim of discovering a relation: if we want to control some outcome or want to predict the value of some unknown variable or merely to understand a relationship.

The second question, about the usefulness and persuasiveness of the line, also depends on the relative sizes of the modeled part of Y and the error. Returning to the notation introduced, this means the relative sizes of the predictable part of Y, $\beta_0 + \beta_1 X$, versus the size of ε. As epsilon gets larger relative to the predictable part, the usefulness of the model declines.

The second question, about how to tell how well a line describes data, can be answered directly with statistics, and it can be answered for quite general cases.

How can we try to guard against seeing relationships where, in fact, none actually exist?

To answer this question we must think like statisticians, do mental handstands, look at the world upside-down.

Remember, the first step in "thinking like a statistician" is to ask, What if there were actually no relationship; zero relationship (so $\beta_1 = 0$)? What would we see?

If there were no relationship then Y would be determined just by random error, unrelated to X. But this does not automatically mean that we would estimate a zero slope for the fitted line. In fact we are highly unlikely to ever estimate a slope of exactly zero. We usually assume that the errors are symmetric, i.e. if the actual value of Y is sometimes above and sometimes below the modeled value, without some oddball skew up or down. So even in a case where there is actually a zero relationship between Y and X, we might see a positive or negative slope.

We would hope that these errors in the estimated slope would be small – but, again, "how small is small?"

Let's take another example. Suppose that the true model is Y = 10 + 2X (so $\beta_0 = 10$ and $\beta_1 = 2$). But of course there will be an error; let's consider a case where the error is pretty large. In this case we might see a set of points like this:

true X and Y
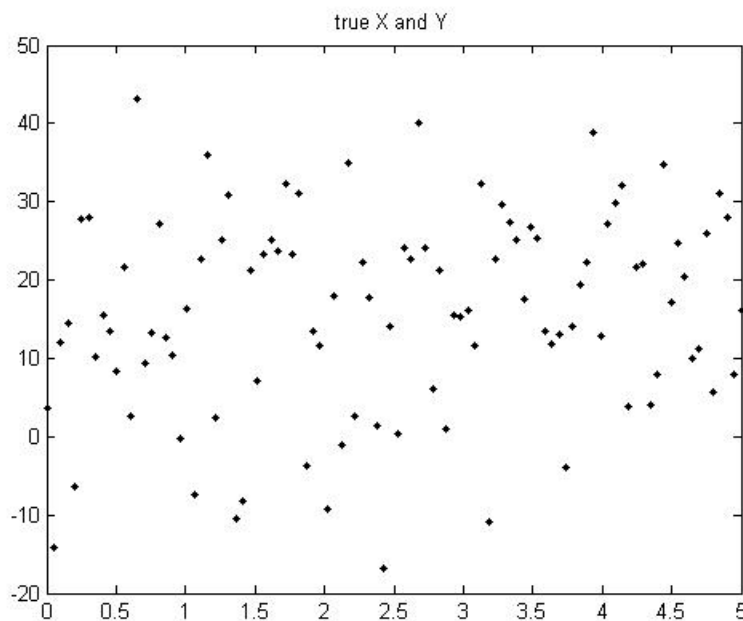
When we estimate the slope for those dots, we would find not 2 but, in this case (for this particular set of errors), 1.61813.

Now we consider a rather strange thing: suppose that there were actually zero relationship between X and Y (so that actually $\beta_1 = 0$). Next suppose that, even though there were actually zero relation, we tried to plot a line and so calculated our estimate of β_1. To give an example, we would have the computer calculate some random numbers for X and Y values, then estimate the slope, and we would find 1.45097. Do it again, and we might get 0.36131. Do it 10,000 times (not so crazy, actually – the computer does it in a couple of seconds), and we'd find the following range of values for the estimated slope:
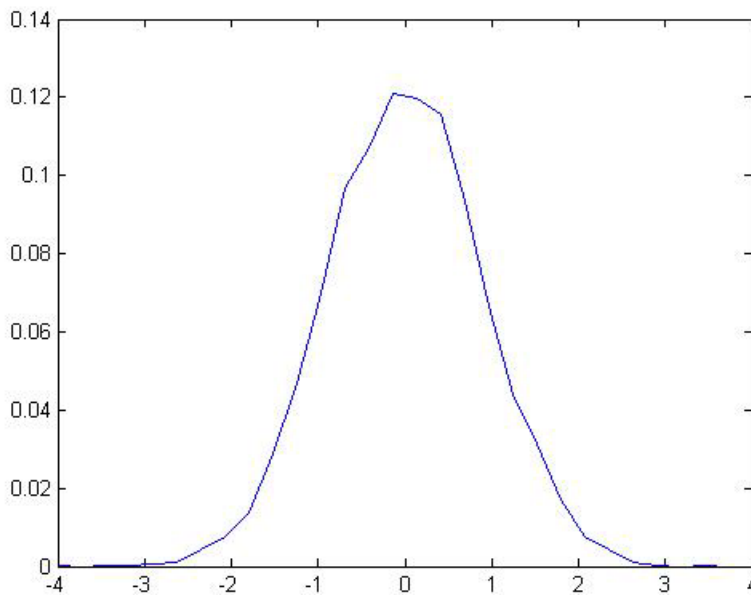
So our estimated slope from the first time, 1.61813, is "pretty far" from zero. How far? The estimated slope is farther than just 659 of those 10,000 tries, which is 6.59%.

So we could say that, if there were actually *no* relationship between X and Y, but we incorrectly estimated a slope, then we'd get something from the range of values shown above. Since we estimated a value of 1.61813, which is farther from zero than just 6.59% if there were actually no relationship, we might say that "there is just a 6.59% chance that X and Y could truly be unrelated but I'd estimate a value of 1.61813."

Now this is a more reasonable measure: "What is the chance that I would see the value, that I've actually got, if there truly were no relationship?" And this percentage chance is relevant and interesting to think about.

This formalization is "hypothesis testing". We have a hypothesis, for example "there is zero relation between X and Y," which we want to test. And we'd like to set down rules for making decisions so that reasonable people can accept a level of evidence as proving that they were wrong. (An example of not accepting evidence: the tobacco companies remain highly skeptical of evidence that there is a relationship between smoking and lung cancer. Despite what most researchers would view as mountains of evidence, the tobacco companies insist that there is some chance that it is all just random. They're right, there is "some chance" – but that chance is, by now, probably something less than 1 in a billion.) Most empirical research uses a value of 5% -- we want to be skeptical enough that there is only a 5% chance that there might really be no relation but we'd see what we saw. So if we went out into the world and did regressions on randomly chosen data, then in 5 out of 100 cases we would think that we had found an actual relation. It's pretty low but we still have to keep in mind that we are fallible, that we will go wrong 5 out of 100 (or 1 in 20) times.

Under some general conditions, the OLS slope coefficient will have a normal distribution -- not a standard normal, though, it doesn't have a mean of zero and a standard deviation of one.

However we can estimate its standard error and then can figure out how likely it is, that the true mean could be zero, but I would still observe that value.

This just takes the observed slope value, call it $\hat{\beta}_1$ (we often put "hats" over the variables to denote that this is the actual observed value), subtract the hypothesized mean of zero, and divide by the standard error:

$$\frac{\hat{\beta}_1 - 0}{se(\beta_1)} = \frac{\hat{\beta}_1}{se(\beta_1)}$$

We call this the "t-statistic". When we have a lot of observations, the t-statistic has approximately a standard normal distribution with zero mean and standard deviation of one.

For the careful students, note that the t-statistic actually has a t-distribution, which has a shape that depends on the number of observations used to construct it (the degrees of freedom). When the number of degrees of freedom is more than 30 (which is almost all of the time), the t-distribution is just about the same as a normal distribution. But for smaller values the t-distribution has fatter tails.

The t-statistic allows us to calculate the probability that, if there were actually a zero relationship, I might actually observe a value as extreme as $\hat{\beta}_1$. By convention we look at distance either above or below zero, so we want to know the probability of seeing a value as far from zero as either $\hat{\beta}_1$ or $-\hat{\beta}_1$. If $\hat{\beta}_1$ were equal to 1, then this would be:
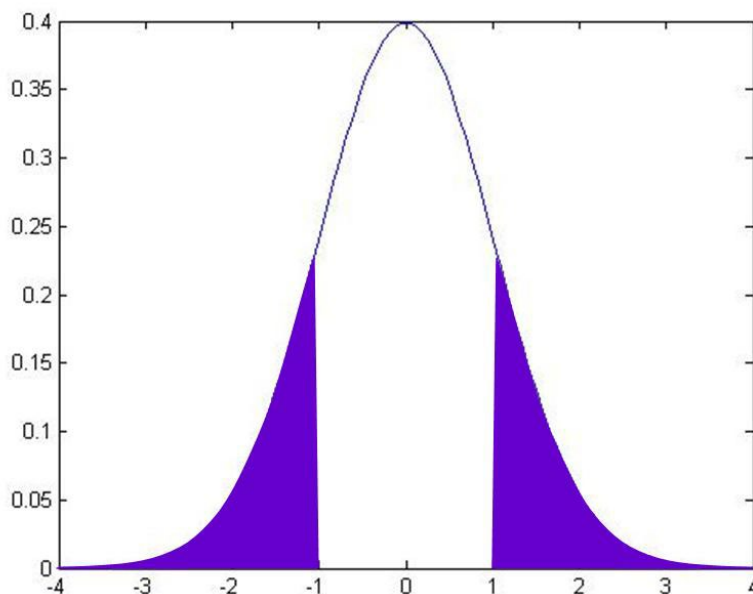
while if $\hat{\beta}_1$ were another value, it would be:
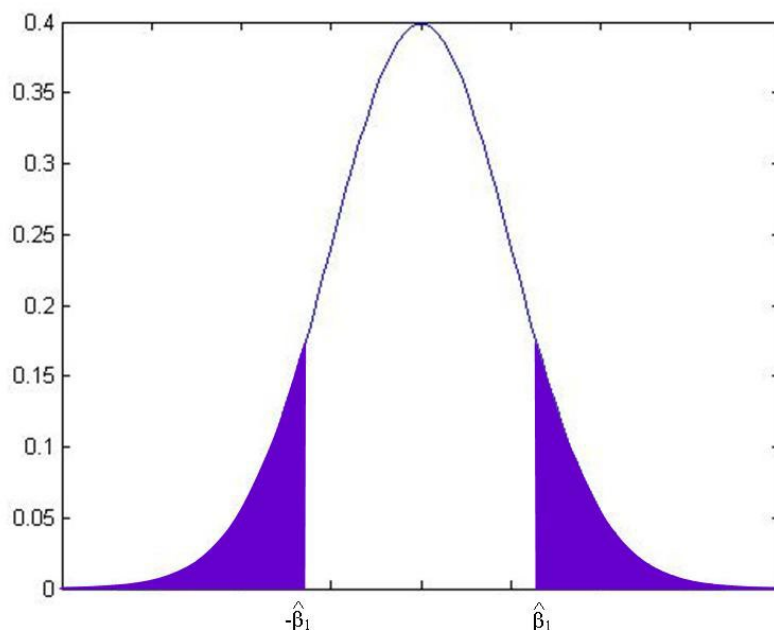
From working on the probabilities under the standard normal, you can calculate these areas for any given value of $\hat{\beta}_1$.

In fact, these probabilities are so often needed, that most computer programs calculate them automatically – they're called "p-values". The p-value gives the probability that the true coefficient could be zero but I would still see a number as extreme as the value actually observed. By convention we refer to slopes with a p-value of 0.05 or less (less than 5%) as "statistically significant".

(We can test if coefficients are different from other values than just zero, but for now that is the most common so we focus on it.)

Confidence Intervals for Regression Estimates

There is another way of looking at statistical significance. We just reviewed the procedure of taking the observed value, subtracting off the mean, dividing by the standard error, and then comparing the calculated t-statistic against a standard normal distribution.

But we could do it backwards, too. We know that the standard normal distribution has some important values in it, for example the values that are so extreme, that there is just a 5% chance that we could observe what we saw, yet the true value were actually zero. This 5% critical value is just below 2, at 1.96. So if we find a t-statistic that is bigger than 1.96 (in absolute value) then the slope would be "statistically significant"; if we find a t-statistic that is smaller than 1.96 (in absolute value) then the slope would not be "statistically significant". We can re-write these statements into values of the slope itself instead of the t-statistic.

We know from above that

$$\frac{\hat{\beta}_1 - 0}{se(\beta_1)} = \frac{\hat{\beta}_1}{se(\beta_1)} = t \text{ ,}$$

and we've just stated that the slope is not statistically significant if:

$$|t| < 1.96.$$

This latter statement is equivalent to:

$$-1.96 < t < 1.96$$

Which we can re-write as:

$$-1.96 < \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} < 1.96$$

Which is equivalent to:

$$-1.96\left(se\left(\hat{\beta}_1\right)\right) < \hat{\beta}_1 < 1.96\left(se\left(\hat{\beta}_1\right)\right)$$

So this gives us a "Confidence Interval" – if we observe a slope within 1.96 standard errors of zero, then the slope is not statistically significant; if we observe a slope farther from zero than 1.96 standard errors, then the slope is statistically significant.

This is called a "95% Confidence Interval" because this shows the range within which the observed values would fall, 95% of the time, if the true value were zero. Different confidence intervals can be calculated with different critical values: a 90% Confidence Interval would need the critical value from the standard normal, so that 90% of the probability is within it (this is 1.64).

Details:

- statistical significance for a univariate regression is the same as overall regression significance – if the slope coefficient estimate is statistically significantly different from zero, then this is equivalent to the statement that the overall regression explains a statistically significant part of the data variation.
- Excel calculates OLS both as regression (from Data Analysis TookPak), as just the slope and intercept coefficients (formula values), and from within a chart
- There are important assumptions about the regression that must hold, if we are to interpret the estimated coefficients as anything other than within-sample descriptors:
    o X completely specifies the causal factors of Y (nothing omitted)
    o X causes Y in a linear manner
    o errors are normally distributed (for small sample test stats)
    o errors have same variance even at different X (homoskedastic not heteroskedastic)
    o errors are independent of each other
- Because OLS squares the residuals, a few oddball observations can have a large impact on the estimated coefficients, so must explore

⚡ Points:

Calculating the OLS Coefficients

The formulas for the OLS coefficients have several different ways of being written. For just one X-variable we can use summation notation (although it's a bit tedious). For more variables the notation gets simpler by using matrix algebra.

The basic problem is to find estimates of β_0 and β_1 to minimize the error in $y_i = \beta_0 + \beta_1 X_i + e_i$.

The OLS coefficients are found from minimizing the sum of squared errors, where each error is defined as $e_i = y_i - \beta_0 - \beta_1 X_i$ so we want to $\min_{\beta_0, \beta_1} \sum_{i=1}^{n} e_i^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i)^2$. If you know basic calculus then you understand that you find the minimum point by taking the derivative with respect to the control variables, so differentiate with respect to β_0 and β_1. After some tedious algebra, find that the minimum value occurs when we use $\hat{\beta}_0$ and $\hat{\beta}_1$, where:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

With some linear algebra, we define the equations as $y = X\beta + e$, where y is a column vector,

$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, e is the same, $e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$, X is a matrix with a first column of ones and then columns of each X

variable, $X = \begin{bmatrix} 1 & x_1^1 & & x_1^k \\ \vdots & \vdots & \ddots & \\ 1 & x_n^1 & & x_n^k \end{bmatrix}$, where there are k+1 columns, and then $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$. The OLS coefficients are

then given as $\hat{\beta} = (X'X)^{-1} X'y$.

But the computer does the calculations so you only need these if you go on to become an econometrician.

To Recap:

- A zero slope for the line is saying that there is no relationship.

- A line has a simple equation, that Y = β_0 + β_1X

- How can we "best" find a value of β?

- We know that the line will not always fit every point, so we need to be a bit more careful and write that our observed Y values, Y_i (i=1, ..., N), are related to the X values, X_i, as: $Y_i = \beta_0 + \beta_1 X_i + u_i$. The u_i term is an error – it represents everything that we haven't yet taken into consideration.

- Suppose that we chose values for β_0 and β_1 that minimized the squared values of the errors. This would mean $\min\limits_{\beta_0,\beta_1} \sum\limits_{i=1}^{N} u_i^2 = \min\limits_{\beta_0,\beta_1} \sum\limits_{i=1}^{N} \left(Y_i - \beta_0 - \beta_1 X_i \right)^2$. This will generally give us unique values of β (as opposed to the eyeball method, where different people can give different answers).

- The β_0 term is the intercept and the β_1 term is the slope, $\dfrac{dY}{dX}$.

- These values of β are the Ordinary Least Squares (OLS) estimates. If the Greek letters denote the true (but unknown) parameters that we're trying to estimate, then denote $\hat{\beta}_0$ and $\hat{\beta}_1$ as our estimators that are based on the particular data. We denote $\hat{Y}_i$ as the predicted value of what we would guess Yi would be, given our estimates of β_0 and β_1, so that $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

- There are formulas that help people calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ (rather than just guessing numbers); these are:

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{N} \left(X_i - \bar{X} \right)\left(Y_i - \bar{Y} \right)}{\sum\limits_{i=1}^{N} \left(X_i - \bar{X} \right)^2} \text{ and}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \text{ so that } \frac{1}{N}\sum\limits_{i=1}^{N} \hat{Y}_i = \bar{Y} \text{ and } \frac{1}{N}\sum\limits_{i=1}^{N} \hat{u}_i = 0$$

Why OLS? It has a variety of desirable properties, if the data being analyzed satisfy some very basic assumptions. Largely because of this (and also because it is quite easy to calculate) it is widely used in many different fields. (The method of least squares was first developed for astronomy.)

- OLS requires some basic assumptions:

  o The conditional distribution of u_i given X_i has a mean of zero. This is a complicated way of saying something very basic: I have no additional information outside of the model, which would allow me to make better guesses. It can also be expressed as implying a zero correlation between X_i and u_i. We will work up to other methods that incorporate additional information. But this is why economists look for "natural experiments" where some X is determined by chance outside the ordinary interrelationships.

  o The X and e are i.i.d. This is often not precisely true; on the other hand it might be roughly right, and it gives us a place to start.

  o X_i and u_i have fourth moments. This is technical and broadly true, whenever the X and Y data have a limit on the amount of variation, although there might be particular circumstances where it is questionable (sometimes in finance).

- These assumptions are costly; what do they buy us? First, if true then the OLS estimates are distributed normally in large samples. Second, it tells us when to be careful.

- Must distinguish between dependent and independent variables (no simultaneity).

- If these are true then the OLS are unbiased and consistent. So $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$. The normal distribution, as the sample gets large, allows us to make hypothesis tests about the values of the betas. In particular, if you look back to the "eyeball" data at the beginning, you will recall that a zero value for the slope, β₁, is important. It implies no relationship between the variables. So we will commonly test the estimated values of β against a null hypothesis that they are zero.

- There are formulas that you can use, for calculating the standard errors of the β estimates, however for now there's no need for you to worry about them. The computer will calculate them. (Also note that the textbook uses a more complicated formula than other texts, which covers more general cases. We'll talk about that later.)

Regression in R

To have R do a linear regression, we use the command "lm()" as for example
```
model1 <- lm(Y ~ X1)
summary(model1)
```

This estimates a linear model of $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ and reports estimates of the intercept and slope coefficients.

Or with the PUMS NY data, create a variable for fraction of income spent on housing then replace the zero-income (thus Inf value) with missing NA values:
```
fraction_housing <- (OWNCOST + RENT)/INCTOT
is.na(fraction_housing) <- is.infinite(fraction_housing)
model2 <- lm(fraction_housing ~ AGE)
summary(model2)
```

Regression Details

We'll often form hypotheses about regression coefficients: t-stats, p-values, and confidence intervals – so that's the same basic process as before. Usually two-sided (rarely one-sided).

We will commonly test if the coefficients 'are significant' – i.e. is there evidence in the data that the coefficient is different from zero? This goes back to our original example where we looked at the difference between the Hong Kong/Singapore stock returns and the US stock returns/interest rate. A zero slope is evidence against any relationship – this shows that the best guess of the value of Y does not depend on current information about the level of X. So coefficient estimates that are statistically indistinguishable from zero are not evidence that the particular X variable is useful in prediction.

A hypothesis test of some statistical estimate uses this estimator (call it $\hat{X}$) and the estimator's standard error (denote it as $se_{\hat{X}}$) to test against some null hypothesis value, X_{null}. To make the hypothesis test, form $Z = \dfrac{\hat{X} - X_{null}}{se_{\hat{X}}}$, and – here is the magic! – under certain conditions this Z will have a Standard Normal distribution (or sometimes, if there are few degrees of freedom, a t-distribution; later in more

advanced stats courses, some other distribution). The magic happens because if Z has a Standard Normal distribution then this allows me to measure if the estimate of X, $\hat{X}$, is very far away from X_{null}. It's generally tough to specify a common unit that allows me to say sensible things about "how big is big?" without some statistical measure. The p-value of the null hypothesis tells me, "If the null hypothesis were actually true, how likely is it that I would see this $\hat{X}$ value?" A low p-value tells me that it's very unlikely that my hypothesis could be true and yet I'd see the observed values, which is evidence against the null hypothesis.

Often the formula, $Z = \dfrac{\hat{X} - X_{null}}{se_{\hat{X}}}$, gets simpler when X_{null} is zero, since it is just $Z' = \dfrac{\hat{X} - 0}{se_{\hat{X}}} = \dfrac{\hat{X}}{se_{\hat{X}}}$, and this is what R prints out in the regression output labeled as "t".

This is in Chapter 5 of Stock & Watson.

We know that the standard normal distribution has some important values in it, for example the values that are so extreme, that there is just a 5% chance that we could observe what we saw, yet the true value were actually zero. This 5% critical value is just below 2, at 1.96. So if we find a t-statistic that is bigger than 1.96 (in absolute value) then the slope would be "statistically significant"; if we find a t-statistic that is smaller than 1.96 (in absolute value) then the slope would not be "statistically significant". We can re-write these statements into values of the slope itself instead of the t-statistic.

We know from above that

$$\frac{\hat{\beta}_1 - 0}{se(\beta_1)} = \frac{\hat{\beta}_1}{se(\beta_1)} = t,$$

and we've just stated that the slope is not statistically significant if:

$$|t| < 1.96.$$

This latter statement is equivalent to:

$$-1.96 < t < 1.96$$

Which we can re-write as:

$$-1.96 < \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} < 1.96$$

Which is equivalent to:

$$-1.96\left(se\left(\hat{\beta}_1\right)\right) < \hat{\beta}_1 < 1.96\left(se\left(\hat{\beta}_1\right)\right)$$

So this gives us a "Confidence Interval" – if we observe a slope within 1.96 standard errors of zero, then the slope is not statistically significant; if we observe a slope farther from zero than 1.96 standard errors, then the slope is statistically significant.

This is called a "95% Confidence Interval" because this shows the range within which the observed values would fall, 95% of the time, if the true value were zero. Different confidence intervals can be

calculated with different critical values: a 90% Confidence Interval would need the critical value from the standard normal, so that 90% of the probability is within it (this is 1.64).

OLS is nothing particularly special. The Gauss-Markov Theorem tells us that OLS is **BLUE**: **B**est **L**inear **U**nbiased **E**stimator (and need to assume homoskedasticity). Sounds good, right? Among the linear unbiased estimators, OLS is "best" (defined as minimizing the squared error). But this is like being the best-looking economist – best within a very small and very particular group is not worth much! Nonlinear estimators may be good in various situations, or we might even consider biased estimators.

If X is a binary dummy variable

Sometimes the variable X is a binary variable, a dummy, D_i, equal to either one or zero (for example, female). So the model is $Y_i = \beta_0 + \beta_1 D_i + u_i$ can be expressed as $Y_i = \begin{cases} \beta_0 + \beta_1 + u_i & if\ D_i = 1 \\ \beta_0 + u_i & if\ D_i = 0 \end{cases}$. So this is just saying that Y has mean $\beta_0 + \beta_1$ in some cases and mean β_0 in other cases. So β_1 is interpreted as the difference in mean between the two groups (those with D=1 and those with D=0). Since it is the difference, it doesn't matter which group is specified as 1 and which is 0 – this just allows measurement of the difference between them.

Other 'tricks' of time trends (& functional form)

•   If the X-variable is just a linear change [for example, (1,2,3,...25) or (1985, 1986,1987,...2010)] then regressing a Y variable on this is equivalent to taking out a linear trend: the errors are the deviations from this trend. Either the X-variable of (1,2,3,...) or (1985,1986,1987,...) gives the same since the slope coefficient estimates dY/dX and in either case dX=1. There is a difference in the intercept term only.

•   If the Y-variable is a log function then the regression is interpreted as explaining percent deviations (since derivative of lnY = dY/Y, the percent change). (So what would a linear trend on a logarithmic form look like?)

•   If both Y and X are logs then can interpret the coefficient as the elasticity.

•   examine errors to check functional form – e.g. height as a function of age works well for age < 12 but then breaks down

•   plots of X vs. both Y and predicted-Y are useful, as are plots of X vs. error.

In addition to the standard errors of the slope and intercept estimators, the regression line itself has a standard error.

A commonly overall assessment of the quality of the regression is the R^2 (displayed by many statistical programs). This is the fraction of the variance in Y that is explained by the model so $0 \leq R^2 \leq 1$. Bigger is usually better, although different models have different expectations (i.e. it's graded on a curve).

Statistical significance for a univariate regression is the same as overall regression significance – if the slope coefficient estimate is statistically significantly different from zero, then this is equivalent to the statement that the overall regression explains a statistically significant part of the data variation.

- Excel calculates OLS both as regression (from Data Analysis TookPak), as just the slope and intercept coefficients (formula values), and from within a chart

Multiple Regression – more than one X variable

Regressing just one variable on another can be helpful and useful (and provides a great graphical intuition) but it doesn't get us very far.

Suppose we wanted to look at a modern version of the classic Engel curve study: what fraction of expenditure goes to food? With the CEX data, we can define the fraction spent on food,

fraction_food <- FOODPQ/TOTEXPPQ

fraction_food[is.infinite(fraction_food)] <- NA

fraction_food[fraction_food<0] <- NA # 1 reported negative total expenditure?!

There are probably lots of factors driving this variation. For example, people who label themselves as white, African-American, Asian, Native American, other race, and Hispanic have different average expenditures. Households where the reference person is African-American spend an average of 19.6% on food, Asians spend 17.5% on food, Native Americans spend 19.1%, other races spend 20.8%, whites spend 17.8%, and Hispanics (who may be of any race) spend 21.7%. (I will leave it as an exercise to determine if these are statistically significantly different.)

There are other differences: people in their 20s average 20.13%, in their 30s spend 18.1%, in their 40s it's down to 17.6%, in 50s 16.8%, then people 60 and up spend 17.8% (somewhat larger). There is a strong relationship with education as well: from those without a high-school diploma who spend 22.9% to those with an advanced degree who spend just 14.4% - suggesting that total income probably is important as well.

So how can we keep all of these different factors straight?

Multiple Regression in R

Chapter 3 of *Applied Econometrics in R* by Kleiber and Zeileis is terrific – gives an enormous amount of detail for how to do lots of different things! Most of this section of notes is based on material from that book. They created a package, *AER*, Applied Econometrics in R, which has lots of useful functions – so load that in.

From the standpoint of just using R, there is little difference for the user between a univariate and multivariate linear regression. Again use "lm()" but then add a bunch of variables to the model specification, so "Y ~ X1 + X2 + X3".

In formulas, model has k explanatory variables for each of $i = (1, 2, \ldots n)$ observations (must have n > k)

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

Each coefficient estimate, notated as $\hat{\beta}_j$, has standardized distribution as t with (n – k) degrees of freedom.

Each coefficient represents the amount by which the y would be expected to change, for a small change in the particular x-variable (i.e. $\beta_j = \dfrac{\partial y}{\partial x_j}$).

Note that you must be a bit careful specifying the variables. Educational attainment might be coded with a bunch of numbers from 31 to 46 but these numbers have no inherent meaning. So too race, geography, industry, and occupation. If a person graduates high school then their grade coding changes from 38 to 39 but this must be coded with a dummy variable. If a person moves from New York to North Dakota then this increases their state code from 36 to 38; this is not the same change as would occur for someone moving from North Dakota to Oklahoma (40) nor is it half of the change as would occur for someone moving from New York to North Carolina (37). Each state needs a dummy variable. These X-variables are not continuous.

A multivariate regression can control for all of the different changes to focus on each item individually. So we might model a household's fraction of expenditure on food as a function of their age, family size, gender of the reference person, race/ethnicity, educational level (high school diploma, some college but no degree, Associate's, a 4-year degree, or advanced degree), if they're married or divorced/widowed/separated, and so forth.

These results are:

```
Model3 <- lm(fraction_food ~ AGE_REF + FAM_SIZE + female + AfAm + Asian +
race_oth + Amindian + Hispanic + educ_hs + educ_smcoll + educ_as +
educ_bach + educ_adv)
```

```
Call:
lm(formula = fraction_food ~ AGE_REF + FAM_SIZE + female + AfAm +
    Asian + race_oth + Amindian + Hispanic + educ_hs + educ_smcoll +
    educ_as + educ_bach + educ_adv)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22494 -0.06511 -0.01622  0.04491  0.83229

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.282e-01  6.472e-03  35.251  < 2e-16 ***
AGE_REF     -4.059e-04  7.565e-05  -5.366 8.31e-08 ***
FAM_SIZE    -1.140e-03  8.719e-04  -1.308   0.1911
female      -4.303e-04  2.480e-03  -0.174   0.8622
AfAm         1.931e-02  3.771e-03   5.121 3.12e-07 ***
Asian        7.080e-03  5.812e-03   1.218   0.2232
race_oth     2.686e-02  1.067e-02   2.518   0.0118 *
Amindian     7.390e-03  1.370e-02   0.539   0.5896
Hispanic     3.055e-02  3.904e-03   7.824 5.88e-15 ***
educ_hs     -2.076e-02  3.995e-03  -5.197 2.08e-07 ***
educ_smcoll -3.235e-02  4.237e-03  -7.634 2.58e-14 ***
educ_as     -4.113e-02  5.024e-03  -8.187 3.17e-16 ***
educ_bach   -5.292e-02  4.306e-03 -12.292  < 2e-16 ***
educ_adv    -6.481e-02  5.296e-03 -12.238  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1018 on 6823 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.05925,  Adjusted R-squared:  0.05746
F-statistic: 33.06 on 13 and 6823 DF,  p-value: < 2.2e-16
```

Take the output a piece at a time. First it confirms what model you had called (useful when you go back later, after you've run lots of regressions). Next it gives a summary of the residuals,

$$\varepsilon_i = y_i - \hat{y} = y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_{1,i} + \widehat{\beta_2} x_{2,i} + \cdots + \widehat{\beta_k} x_{k,i} \right)$$

These can be called at any point with "`residuals(model3)`" so the output is simply from "`summary(residuals(model3))`". The mean is not reported here since the model constrains the mean of the residuals to zero. The fitted values, $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_{1,i} + \widehat{\beta_2} x_{2,i} + \cdots + \widehat{\beta_k} x_{k,i}$, can be called as `fitted.values(model3)`. You can plot these.

Then R reports the coefficients, standard errors, t-statistics, and p-values for each term in the model. The coefficients and standard errors are calculated by the estimation routine. The t-statistic is the ratio of the coefficient estimate divided by the standard error, $t = \dfrac{\hat{\beta}}{se(\hat{\beta})}$. The p-value is the area in the tails of a t-distribution (with degrees of freedom as shown on bottom line, here "6823 DF") beyond the t-statistic. The command, "`coefficients(model3)`", accesses the coefficient values.

At the bottom of the R summary it shows the R-squared, the standard error of the residual (which is basically the same as `sd(residuals(model3))`), and the F-statistic, which is another measure of how well the model fits.

Residuals are often used in analyses of productivity. Suppose I am analyzing a chain's stores to try to figure out which are managed best. I know that there are many reasons for variation in revenues and cost so I can get data on those: how many workers are there and their pay, the location of the store relative to traffic, the rent paid, any sales or promotions going on, etc. A regression on all of those factors delivers an estimate, $\hat{y}$, of what profit would have been expected, given external factors. Then the difference represents the unexplained or residual amount of variation: some stores would have been expected to be profitable and are indeed; some are not living up to potential; some would not have been expected to do so well but something is going on so they're doing much better than expected. But in general it's tricky to assign a name to the residual – unless that name is "ignorance."

You should be able to calculate the t-statistic and p-value from the coefficient estimates and standard errors by yourself (the next homework will give you some chances to practice that).

You should also be able to calculate confidence intervals, although R can do that for you as well, with for example, `confint(model3,level = 0.95)`.

R will also produce lots of plots, simply with `plot(model3)`, which gives lots of plots in sequence – you can pick off particular ones with `plot(model3, which = 3)` that will give the 3rd plot. (The plots indicate that this might not be a great model.)

You can get an Analysis of Variance (ANOVA) with `anova(model3)`. For now don't worry about the details of the output except to the final row of figures, labeled "`Residuals`". This gives one of the most important bits of information about the model: how big are the residuals? Remember that's the whole point of the OLS estimator – it minimizes the (squared) residuals. So this gives you the value of the sum of squared residuals.

We often want to know particular predictions, for example we might want to know what the model would predict is the fraction of expenditure for a 30-year-old female, without anyone else in the household,

who is African-American and has a bachelor's degree. To do this in R, we would first create the data frame then use the predict command:

```
to_be_predicted <- data.frame(AGE_REF = 30, FAM_SIZE = 1, female = 1,
                              AfAm = 1, Asian = 0, race_oth = 0, Amindian =0,
                              Hispanic = 0, educ_hs = 0, educ_smcoll = 0,
                              educ_as = 0, educ_bach = 1, educ_adv = 0)
predict(model3, newdata = to_be_predicted, interval = "confidence")
```

There is a final detail, that we use $interval = "confidence"$ if the x-values to be predicted are inside the values estimated, and $interval = "prediction"$ if the x-values are outside.

(yeah, these notes start to get skimpy – you might want to, you know, actually go to class!)

CPS Data

We have been using various data sets; today we'll use another well-known data set, the Current Population Survey. This dataset has over 200,000 people; it is the basis for the US unemployment statistics. The Bureau of Labor Statistics (BLS) and Census work together to put together and maintain this data; every March a new group of people rotates in while the old group (who answered questions for the past year) rotate out. It is all publicly available: if some wacko thinks the government is fudging the unemployment statistics, they can go and re-calculate everything on their own to check. I've put the data file, `cps_mar2013.RData`, onto the web along with `cps_mar2013_initial_recoding.r` (which you need not run) that created the file from the original data (and has the details of the coding) as well as `cps_1.R`.

Can run a basic linear regression to find what are principal determinants of wage/salary levels (looking at a subset of prime-age, fulltime, year-round workers):
```
model1 <- lm(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian + race_oth
+ Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach + educ_adv +
married + divwidsep + union_m + veteran + immigrant + immig2gen)
     coeftest(model1)
```

This gives an estimate of how important are various educational qualifications.

Statistical Significance

Statistical significance of coefficient estimates is the same when we look at individual coefficients but more complicated for multiple coefficients: we can ask whether a group of variables are jointly significant, which takes a more complicated test. We can even ask if all of the slope coefficients together are statistically significant.

For a univariate regression, if the single slope coefficient is statistically significant then the overall regression is as well (the F statistic is the square of the t-stat in that case).

The difference between the overall regression fit and the significance of any particular estimate is that a hypothesis test of one particular coefficient tests if that parameter is zero; is $\beta_i = 0$? This uses the t-statistic $t = \dfrac{\hat{\beta}}{se(\hat{\beta})}$ and compares it to a t distribution. The test of the regression significance tests if ALL of the slope coefficients are simultaneously zero; if $\beta_1 = \beta_2 = \beta_3 = ... = \beta_K = 0$. The latter is much more restrictive. (See Chapter 7 of Stock & Watson.)

It is often sensible to make joint tests of regression coefficients, for example with a group of dummy variables. If we have a set of dummies for education levels, it is strange to think of omitting just one or two; it is more reasonable to ask whether education measures (overall) are statistically significant. We might also want to know if individual coefficients are equal to each other (e.g. to ask if going to college, without getting any degree, is really different from the estimate for just a high school diploma.

To do this in R, there is a package, *linearHypothesis* (part of the package, *car*, Companion to Applied Regression, which is auto-loaded by *AER* package). But the commands shouldn't obscure the simple basic point: we evaluate variables based on how well they fit in the model.

To consider the question of whether a set of variables is statistically significant, we basically are just looking at how big is the error (the Sum of Squared Errors) with and without those variables. In general adding more variables to the model can never make the errors bigger (can never increase the Sum of Squared Errors) – basically this is a statement that the Marginal Benefit of more variables can never be negative. But profit maximization requires that we balance Marginal Benefit against Marginal Cost – what is the marginal cost of adding more variables? Statistical significance is one measure of profitability in this sense.

If adding new predictors makes the error "a lot" smaller, then those predictors are jointly statistically significant. The essence of statistical testing is just finding a good metric for "a lot".

Note that we can only properly make comparisons within models – it doesn't make much sense to look across models. If I have a model of the fraction of income spent on food, and another model of the level of income, it is difficulty to sensibly pose a question like, "in which model is education more important?" It would be like asking who scored more points per game, Shaq or Jeter? – you can ask the question but it's difficult to interpret in a sensible way.

But within a model we can make comparisons and many of them come down to asking, how much smaller are the errors? (Did the Sum of Squared Errors fall by a lot?) Sometimes it is easiest to just estimate the model twice, with or without the variables of interest, and look at how much the Sum of Squared Errors (from ANOVA in R) fell. But once you get some experience, you'll appreciate *linearHypothesis*.

Finally note that "statistically significant" is different from "important". Suppose you have some Y-values ranging from 100 – 1000, but you notice that a particular X value is associated with the first decimal value. When X has one value, the first decimal is .2; when X has another value the first decimal is 0.7. There are a lot of reasons that could be the case. This could be an interesting pattern and this could tell us subtle things about the world. But a 0.5 difference, among values ranging over 3 digits, is really tiny! A hypothesis of statistical significance could duly tell you that the X-value is significant (it is a good indicator of whether the outcome is yyy.2 or yyy.7). But depending on the question you're asking, that could be unimportant.

Why do we always leave out a dummy variable? Multicollinearity. (See Chapter 6 of Stock & Watson.)

- OLS basic assumptions:

  o The conditional distribution of u_i given X_i has a mean of zero. This is a complicated way of saying something very basic: I have no additional information outside of the model, which would allow me to make better guesses. It can also be expressed as implying a zero correlation between X_i and u_i. We will work up to other methods that incorporate additional information.

  o The X and errors are i.i.d. This is often not precisely true; on the other hand it might be roughly right, and it gives us a place to start.

  o X and errors don't have values that are "too extreme." This is technical (about existence of fourth moments) and broadly true, whenever the X and Y data have a limit on the amount of variation, although there might be particular circumstances where it is questionable (sometimes in finance).

- So if these are true then the OLS are unbiased and consistent. So $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$. The normal distribution, as the sample gets large, allows us to make hypothesis tests about the values of the betas. In particular, if you look back to the "eyeball" data at the beginning, you

will recall that a zero value for the slope, β_1, is important. It implies no relationship between the variables. So we will commonly test the estimated values of β against a null hypothesis that they are zero.

Factors in R

R has a shortcut for lots of dummy variables – some variables are labeled as factors. You might wonder why I hadn't been using them earlier, but that's because R makes it too easy. You can forget what they mean and just trust in the magic. But as Stevie Wonder sang,

When you believe in things that you don't understand
Then you suffer
Superstition ain't the way
      https://youtu.be/oCFuCYNx-1g

So I don't want you to suffer … or at least I want to tradeoff to do more suffering early so that you will suffer less in the long term (the optimality of that tradeoff obviously depends on your individual intertemporal discount rate).

For instance, you could define a set of dummy variables,

```
educ_hs, educ_smcoll, educ_as, educ_bach, educ_adv
```

Or you could define,

```
educ_indx <- as.factor(educ_nohs + 2*educ_hs + 3*educ_smcoll +
     4*educ_as + 5*educ_bach + 6*educ_adv)
levels(educ_indx)[1] <- "No HS"
levels(educ_indx)[2] <- "HS"
levels(educ_indx)[3] <- "Some Coll"
levels(educ_indx)[4] <- "AS"
levels(educ_indx)[5] <- "Bach"
levels(educ_indx)[6] <- "Adv Deg"
levels(educ_indx)
```

In this case the regression specifications,

```
lm(WSAL_VAL ~ Age + educ_hs + educ_smcoll + educ_as + educ_bach + educ_adv)
```

or

```
lm(WSAL_VAL ~ Age + educ_indx)
```

would give the same results. (Try it!) Note that you have to make sure that R knows (either it guessed when the data was loaded or you explicitly used the "`as.factor()`" command) that the variable is a factor since otherwise it will treat it as a continuous variable. You can use "`is.factor()`" to get info about whether R currently thinks that a particular variable is a factor. R will automatically drop one of the dummy variables and will stack the output in the way it thinks makes sense. But sometimes there needs to be wrestling between what you want and what it wants to give you – so you have to know what is going on underneath and don't just rely on the R magic to do it. That's superstition.

Heteroskedasticity-consistent errors

You can choose to use heteroskedasticity-consistent errors as in the textbook.

The Stock and Watson textbook uses heteroskedasticity-consistent errors (sometimes called Eicker-Huber-White errors, after the various authors who figured out how to calculate them). Later you can additionally specify heteroskedasticity- and autocorrelation-consistent (HAC) errors, sometimes called Newey-West.

In linear regression these don't change the coefficient estimates but just the standard errors of those estimators. (Which is not true for nonlinear cases, which we'll be discussing later.)

Heteroskedasticity-Consistent Errors in R

These are HCerrors, in the "sandwich" package, which depends on "zoo" package; probably the easiest implementation is via the "lmtest" package. So install those 3,

```
library("zoo")
library("sandwich")
library("lmtest")
```

For heteroskedasticity-consistent errors, use the coeftest() function but add the command, vcovHC. So from example of CPS data, use:

```
coeftest(model1,vcovHC)
```

The command `coeftest` will do a variety of coefficient tests; if you don't play with the defaults, you get the same standard errors as in the summary. If you use `vcovHC`, you get the heteroskedasticity-consistent standard errors. (Econometricians have worked their little butts off, coming up with variations on these, so there are HC0 through HC5 just in this package, don't worry for now about which one to use.)

If you compare the two sets of output, you should notice that the actual coefficient estimates are unchanged – it's the estimated standard errors that change. Then those changes propagate through, so the t-statistics and p-values also change. There is no generic result for whether the estimated standard errors are always bigger or smaller and even in the output from this simple case it goes both ways. However the standard errors often tend to be bigger with the heterskedasticity correction (which means that – test yourself! – the t-statistics are ___ [*bigger or smaller in absolute value?*] and p-values are ___ [*bigger or smaller?*]).

```
# cps_1.R
# looking at CPS 2013 data
# uses file from dataferret download of CPS March 2013 supplement,
downloaded June 12 2014
# accompanying lecture notes for KFoster class ECO B2000

rm(list = ls(all = TRUE))
setwd("C:\\Users\\Kevin\\Documents\\CCNY\\data for classes\\CPS_Mar2013")
load("cps_mar2013.RData")

attach(dat_CPSMar2013)
# use prime-age,fulltime, yearround workers
use_varb <- (Age >= 25) & (Age <= 55) & work_fullt & work_50wks
dat_use <- subset(dat_CPSMar2013,use_varb) # 47,550 out of 202,634 obs

detach(dat_CPSMar2013)

attach(dat_use) # just prime-age,fulltime, yearround workers

# always a good idea to get basic stats of all of the variables in your
regression to see if they make sense
summary(WSAL_VAL)
summary(Age)
summary(female)
summary(AfAm)
summary(Asian)
summary(Amindian)
summary(race_oth)
summary(Hispanic)
summary(educ_hs)
summary(educ_smcoll)
summary(educ_as)
summary(educ_bach)
summary(educ_adv)
summary(married)
summary(divwidsep)
summary(union_m)
summary(veteran)
summary(immigrant)
summary(immig2gen)

model1 <- lm(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian + race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen)

summary(model1)
coeftest(model1)
#sometimes log form is preferred
# dat_noZeroWage <- subset(dat_use,(WSAL_VAL > 0))
# model1a <- lm(log(WSAL_VAL) ~ Age + female + AfAm + Asian + Amindian +
race_oth
#              + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
#              + married + divwidsep + union_m + veteran + immigrant +
immig2gen, data = dat_noZeroWage)
# detach(dat_use)
# attach(dat_noZeroWage)
# log(mean(WSAL_VAL))
# mean(log(WSAL_VAL))
# detach(dat_noZeroWage)
# attach(dat_use)
```

```
# ^^ yes there are more elegant ways to do that, avoiding attach/detach -
find them!

# for heteroskedasticity consistent errors
require(sandwich)
require(lmtest)

coeftest(model1,vcovHC)

# jam nonlinear into linear regression
model2 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + AfAm + Asian + Amindian +
race_oth
            + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
            + married + divwidsep + union_m + veteran + immigrant +
immig2gen)

model3 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + I(female*Age) +
I(female*(Age^2)) + AfAm + Asian + Amindian + race_oth
            + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
            + married + divwidsep + union_m + veteran + immigrant +
immig2gen)
# could do this with "update" function instead
summary(model2)
summary(model3)
# the ANOVA function is flexible - can compare nested models
anova(model1,model2,model3)

# Applied Econometrics in R suggests also spline and kernel estimators, we
might get to that later

# subset in order to plot...
NNobs <- length(WSAL_VAL)
set.seed(12345) # just so you can replicate and get same "random" choices
graph_obs <- (runif(NNobs) < 0.1) # so something like 4000 obs
dat_graph <-subset(dat_use,graph_obs)

plot(WSAL_VAL ~ jitter(Age, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5,
alpha = 0.02), data = dat_graph)
# ^^ that looks like crap since Wages are soooooooo skew!  So try to find
ylim = c(0, ??)
plot(WSAL_VAL ~ jitter(Age, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5,
alpha = 0.02), ylim = c(0,150000), data = dat_graph)

# to plot the predicted values might want to do something like,
lines(fitted.values(model2) ~ Age)
# but that will plot ALLLLL the values, which is 4500 too many and looks
awful
# so back to this,
to_be_predicted2 <- data.frame(Age = 25:55, female = 1, AfAm = 0, Asian = 0,
Amindian = 1, race_oth = 1,
                                Hispanic = 1, educ_hs = 0, educ_smcoll = 0,
educ_as = 0, educ_bach = 1, educ_adv = 0,
                                married = 0, divwidsep =0, union_m = 0,
veteran = 0, immigrant = 0, immig2gen = 1)
to_be_predicted2$yhat <- predict(model2, newdata = to_be_predicted2)

lines(yhat ~ Age, data = to_be_predicted2)

# now compare model3
to_be_predicted3m <- data.frame(Age = 25:55, female = 0, AfAm = 0, Asian =
0, Amindian = 1, race_oth = 1,
```

```
                                    Hispanic = 1, educ_hs = 0, educ_smcoll = 0,
educ_as = 0, educ_bach = 1, educ_adv = 0,
                                    married = 0, divwidsep =0, union_m = 0,
veteran = 0, immigrant = 0, immig2gen = 1)
to_be_predicted3m$yhat <- predict(model3, newdata = to_be_predicted3m)

to_be_predicted3f <- data.frame(Age = 25:55, female = 1, AfAm = 0, Asian =
0, Amindian = 1, race_oth = 1,
                                    Hispanic = 1, educ_hs = 0, educ_smcoll = 0,
educ_as = 0, educ_bach = 1, educ_adv = 0,
                                    married = 0, divwidsep =0, union_m = 0,
veteran = 0, immigrant = 0, immig2gen = 1)
to_be_predicted3f$yhat <- predict(model3, newdata = to_be_predicted3f)

plot(WSAL_VAL ~ jitter(Age, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5,
alpha = 0.02), ylim = c(0,150000), xlab = "Age", data = dat_graph)
lines(yhat ~ Age, data = to_be_predicted3f)
lines(yhat ~ Age, data = to_be_predicted3m, lty = 2)
legend("topleft", c("male", "female"), lty = c(2,1), bty = "n")

det_ind <- as.factor(A_DTIND)
det_occ <- as.factor(A_DTOCC)

model4 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + AfAm + Asian + Amindian +
race_oth
            + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
            + married + divwidsep + union_m + veteran + immigrant +
immig2gen
            + det_ind + det_occ)
summary(model4)

# and always remember this part...
detach(dat_use)
```

Nonlinear Regression

(more properly, **How to Jam Nonlinearities into a Linear Regression**)

- X, X^2, X^3, ... X^r

- $\ln(X)$, $\ln(Y)$, both $\ln(Y)$ & $\ln(X)$

- dummy variables

- interactions of dummies

- interactions of dummy/continuous

- interactions of continuous variables

There are many examples of, and reasons for, nonlinearity. In fact we can think that the most general case is nonlinearity and a linear functional form is just a convenient simplification which is sometimes useful. But sometimes the simplification has a high price. For example, my kids believed that age and height are closely related – which is true for their sample (i.e. mostly kids of a young age, for whom there is a tight relationship, plus 2 parents who are aged and tall). If my sample were all children then that might be a decent simplification; if my sample were adults then that's lousy.

The usual justification for a linear regression is that, for any differentiable function, the Taylor Theorem delivers a linear function as being a close approximation – but this is only within a neighborhood. We need to work to get a good approximation. This might be the case for instance for macro models – we have decent econometric predictions of what happens if the Fed bumps rates from 4% to 4.25%, because we have a lot of similar cases. But what happens when rates stay at 0%? We don't have nearly as much evidence (certainly before 2008, very little). So non-linearities were more important.

Nonlinear terms

We can return to our regression using CPS data. First, we might want to ask why our regression is linear. This is mostly convenience, and we can easily add non-linear terms such as Age^2, if we think that the typical age/wage profile looks like this:
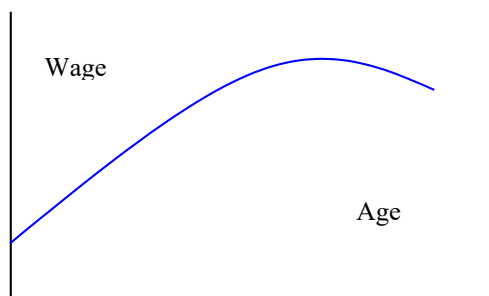
So the regression would be:

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \ldots + \varepsilon_i$$

(where the term "..." indicates "other stuff" that should be in the regression).

As we remember from calculus,

$$\frac{dWage}{dAge} = \beta_1 + \beta_2 \cdot 2 \cdot Age$$

so that the extra "boost" in wage from another birthday might fall as the person gets older, and even turn negative if the estimate of $\beta_2 < 0$ (a bit of algebra can solve for the top of the hill by finding the Age that sets $\frac{dWage}{dAge} = 0$).

We can add higher-order effects as well. Some labor econometricians argue for including Age³ and Age⁴ terms, which can trace out some complicated wage/age profiles. However we need to be careful of "overfitting" – adding more explanatory variables will never lower the R².

To show this in R, I will do a lot of plots – details in `cps_1.R`. (below)
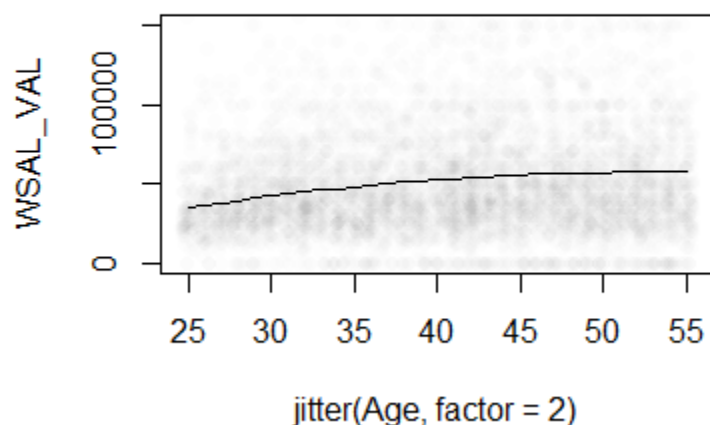
Logarithms

Similarly can specify X or Y as ln(X) and/or ln(Y).

(You also need to figure out how to work with observations where Y=0 since ln(0) doesn't give good results. Dropping those observations might be OK or might not, it depends.)

But we've got to be careful: remember from math (or theory of insurance from Intermediate Micro) that E[ln(Y)] **IS NOT EQUAL TO** ln(E[Y]) ! In cases where we're regressing on wages, this means that the log of the average wage is not equal to the average log wage.

(Try it. Go ahead, I'll wait.)

When both X and Y are measured in logs then the coefficients have an easy economic interpretation. Recall from calculus that with $y = \ln(x)$ and $\frac{dy}{dx} = \frac{1}{x}$, so $dy = \frac{dx}{x} = \%\Delta x$ -- our usual friend, the percent change. So in a regression where both X and Y are in logarithms, then $\beta_j = \frac{\Delta y}{\Delta x} = \frac{\%\Delta y}{\%\Delta x}$ is the elasticity of Y with respect to X.

Also, if Y is in logs and D is a dummy variable, then the coefficient on the dummy variable is just the percent change when D switches from zero to one.

So the choice of whether to specify Y as levels or logs is equivalent to asking whether dummy variables are better specified as having a constant level effect (i.e. women make $10,000 less than men) or having a percent change effect (women make 25% less than men). As usual there may be no general answer that one or the other is always right!

Recall our discussion of dummy variables, that take values of just 0 or 1, which we'll represent as D_i. Since, unlike the continuous variable Age, D takes just two values, it represents a shift of the constant term. So the regression,

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_3 D_i + u_i$$

The equation could be also written as

$$Wage_i = \begin{cases} \beta_0 + \beta_1 Age_i + u_i & for\ D = 0 \\ \beta_0 + \beta_3 + \beta_1 Age_i + u_i & for\ D = 1 \end{cases}.$$

These show that people with D=0 have intercept of just β_0, while those with D=1 have intercept equal to $\beta_0 + \beta_3$. Graphically, this is:
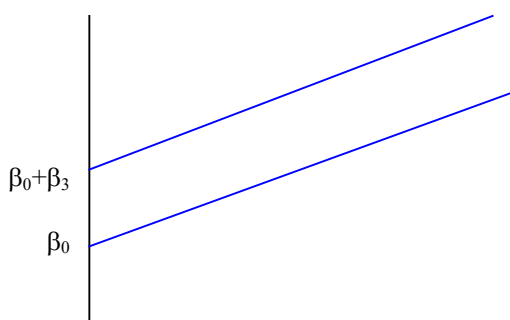
We need not assume that the β_3 term is positive – if it were negative, it would just shift the line downward. We *do* however assume that the *rate* at which age increases wages is the same for both genders – the lines are parallel.

Dummy Variables Interacting with Other Explanatory Variables

The assumption about parallel lines with the same slopes can be modified by adding interaction terms: define a variable as the product of the dummy times age, so the regression is

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_3 D_i + \beta_4 D_i Age_i + u_i$$

or

$$Wage_i = (\beta_0 + \beta_3 D_i) + (\beta_1 + \beta_4 D_i) Age_i + u_i$$

or

$$Wage_i = \begin{cases} \beta_0 + \beta_1 Age_i + u_i & for\ D = 0 \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_4) Age_i + u_i & for\ D = 1 \end{cases}$$

so that, for those with D=0, as before $\dfrac{\Delta Wage}{\Delta Age}$ =β_1 but for those with D=1, $\dfrac{\Delta Wage}{\Delta Age} = \beta_1 + \beta_4$.

Graphically,
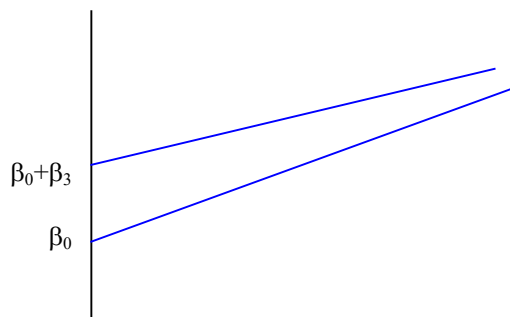
so now the intercepts and slopes are different.

So we might wonder if men and women have a similar wage-age profile. We could fit a number of possible specifications that are variations of our basic model that wage depends on age and age-squared. The first possible variation is simply that:

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 D_i + u_i ,$$

which allows the wage profile lines to have different intercept-values but otherwise to be parallel (the same hump point where wages have their maximum value), as shown by this graph:
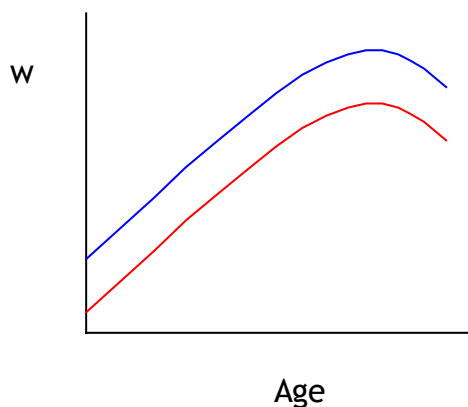
The next variation would be to allow the lines to have different slopes as well as different intercepts:

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2$$
$$+ \beta_3 D_i + \beta_4 D_i Age_i + \beta_5 D_i Age_i^2 + u_i$$

which allows the two groups to have different-shaped wage-age profiles, as in this graph:
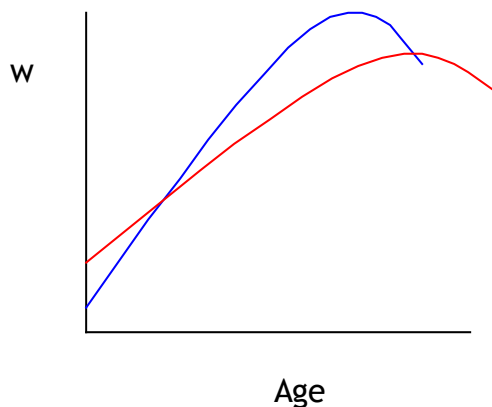
(The wage-age profiles might intersect or they might not – it depends on the sample data.)

We can look at this alternately, that for those with D=0,

$$wage = \beta_0 + \beta_1 Age + \beta_2 Age^2$$
$$\frac{dWage}{dAge} = \beta_1 + 2\beta_2 Age$$

so the extreme value of Age (where $\frac{dWage}{dAge} = 0$) is $\frac{-\beta_1}{2\beta_2}$.

While for those with D=1,

$$Wage = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 + \beta_4 Age + \beta_5 Age^2$$
$$= (\beta_0 + \beta_3) + (\beta_1 + \beta_4) Age + (\beta_2 + \beta_5) Age^2$$
$$\frac{dWage}{dAge} = (\beta_1 + \beta_4) + 2(\beta_2 + \beta_5) Age$$

so the extreme value of Age (where $\frac{dWage}{dAge} = 0$) is $\frac{-(\beta_1 + \beta_3)}{2(\beta_2 + \beta_4)}$. Or write the general value, for both cases, as $\frac{-(\beta_1 + \beta_3 D)}{2(\beta_2 + \beta_4 D)}$ where D is 0 or 1.

This specification, with a dummy variable multiplying each term: the constant and all the explanatory variables, is equivalent to running two separate regressions: one for men and one for women:

$$D = 0$$
$$Wage_i = \beta_0^{male} + \beta_1^{male} Age_i + \beta_2^{male} Age_i^2 + u_i$$
$$D = 1$$
$$Wage_i = \beta_0^{female} + \beta_1^{female} Age_i + \beta_2^{female} Age_i^2 + e_i$$

Where the new coefficients are related to the old by the identities: $\beta_0^{female} = \beta_0 + \beta_3$, $\beta_1^{female} = \beta_1 + \beta_4$, and $\beta_2^{female} = \beta_2 + \beta_5$. Sometimes breaking up the regressions is easier, if there are large datasets and many interactions.

Note that it would be very weird (and difficult to justify) to have an interaction of the dummy with the Age term but not with Age-squared or vice versa. Why would we want to assume that, say, men and women have different linear effects but the same squared effect?

The plot for the CPS data is (code is below):
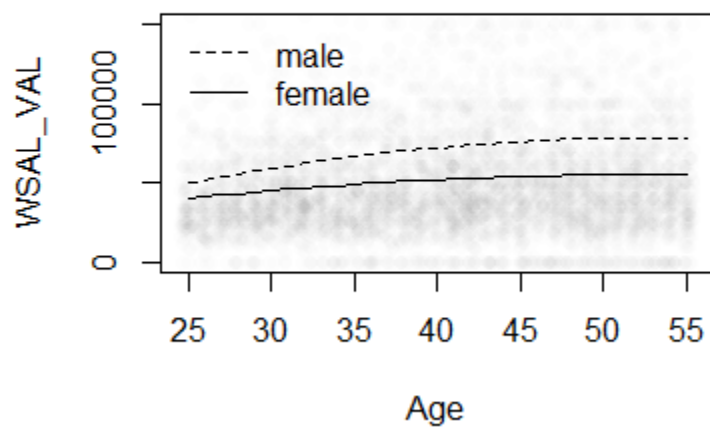
Interactions with R
It is very easy to do interactions with R, maybe too easy so that you can forget what it all means.

That's one reason I'm creating education dummies separately rather than using the R shortcut (creating a single education index of type "`factor`"), so that you can better see what's going on underneath. Once you understand the basics, you can start using the shortcuts.

In a formula, you can do interactions of, say, gender with education with just the ":" operator, so with the factor of "`educ_indx`", include "`female:educ_indx`"

```
model1 <- lm(WSAL_VAL ~ Age + female + educ_indx + female:educ_indx + AfAm
+ Asian + Amindian + race_oth + Hispanic + married + divwidsep + union_m +
veteran + immigrant + immig2gen)

summary(model1)
```

It can be difficult to unpack the meaning all of the interaction terms. The regression creates dummy variables for educational classifications, showing that people with progressively higher educational qualifications get more money. But women get less at each rung: the coefficients on female interacted with education are negative. So for instance a male with an associate's degree is predicted to make about $20,700 more than a male without even a high school diploma, but a woman with an associate's degree gets $8400 less than the man – so her net premium for the associate's degree is (20,700 – 8400) = 12,300.

We can create a table showing the net values, like this (also setting Age = 30),

	Intercept+(Age=30)	HS	Some Coll	AS	Bach	Adv Deg
male	24494	10570	20178	20737	44536	79607
female difference	-6494	-6501	-9045	-8391	-15904	-30213
net	18001	4068	11133	12347	28632	49394

So in equations this says that
$$wage = \beta_0 + \beta_1 Age + \beta_2 Female + \beta_3 EducHS + \beta_4 EducSomeC + \beta_5 EducAS + \beta_6 EducBach + \beta_7 EducAdv + \cdots \{other\ stuff\} +$$

$$\dots + \gamma_1 Female * EducHS + \gamma_2 Female * EducSomeC + \gamma_3 Female * EducAS + \gamma_4 Female * EducBach$$
$$+ \gamma_5 Female * EducAdv + \varepsilon$$

Then the predicted values are, say for a 30-year-old female with an associate's degree,
$$\widehat{wage} = \beta_0 + \beta_1(Age = 30) + \beta_2(Female = 1) + \beta_3(EducHS = 0) + \beta_4(EducSomeC = 0)$$
$$+ \beta_5(EducAS = 1) + \beta_6(EducBach = 0) + \beta_7(EducAdv = 0) + \cdots \{other\ stuff\} +$$
$$\dots + \gamma_1(Female = 1) * (EducHS = 0) + \gamma_2(Female = 1) * (EducSomeC = 0) + \gamma_3(Female = 1)$$
$$* (EducAS = 1) + \gamma_4(Female = 1) * (EducBach = 0) + \gamma_5(Female = 1) * (EducAdv = 0)$$

Which looks ferociously complicated but multiplying by zero drops many of the terms
$$\widehat{wage} = \beta_0 + \beta_1(Age = 30) + \beta_2(Female = 1) + \color{gray}{\beta_3(EducHS = 0)} + \color{gray}{\beta_4(EducSomeC = 0)}$$
$$+ \beta_5(EducAS = 1) + \color{gray}{\beta_6(EducBach = 0)} + \color{gray}{\beta_7(EducAdv = 0)} + \cdots \{other\ stuff\} +$$
$$\dots + \color{gray}{\gamma_1(Female = 1) * (EducHS = 0)} + \color{gray}{\gamma_2(Female = 1) * (EducSomeC = 0)} + \gamma_3(Female = 1)$$
$$* (EducAS = 1) + \color{gray}{\gamma_4(Female = 1) * (EducBach = 0)} + \color{gray}{\gamma_5(Female = 1) * (EducAdv = 0)}$$

> From staring at the wage penalties, you might also conclude that it looks somewhat multiplicative, that the wage penalty for females is around 35%-40% for all of the terms involving college. This might motivate a log specification (which is usually preferred in the literature, I'm just passing over it here in order not to overwhelm with ornamentation).

You might next look at gender/marital status interactions, or education/race/ethnicity interactions – there is no reason you can't do interactions upon interactions. They get complicated but just write out the various interactions in long equation format to help remember what is what. Just don't be a monkey about interpreting and understanding all of the interactions. The limit on how many interactions comes since as you take finer and finer cuts, you're essentially looking at group means where the numbers in each group get smaller and smaller. So you can do state-level factors interacted with gender and education, and probably get a decent estimate of how the wages of women-with-associates-in-NY compares with wages of women-with-associates-in-Cali, but worse estimate of women-with-associates-in-Wyoming or some other empty state where nobody lives. Multi-level models (later) try to deal with this problem.

Testing if All the New Variable Coefficients are Zero

You're wondering how to tell if all of these new interactions are worthwhile. Simple: Hypothesis Testing! There are various formulas, some more complicated, but for the case of homoskedasticity the formula is relatively simple.

Why any formula at all – why not look at the t-tests individually? Because the individual t-tests are asking if each individual coefficient is zero, not if it is zero and others as well are also zero. That would be a stronger test.

To measure how much a group of variables contributes to the regression, we look at the residual values – how much is still unexplained, after the various models? And since this is OL**S**, we look at the **squared** residuals. R outputs the Sum of Squares for the Residuals in the ANOVA. We compare the sum of squares from the two models and see how much it has gone down with the extra variables. A big decrease indicates that the new variables are doing good work. And how do we know, how big is "big"? Compare it to some given distribution, in this case the F distribution. Basically we look at the percent change in the sum of squares, so something like:

$$F \approx \frac{SSR_0 - SSR_1}{SSR_1}$$

with the wavy equals sign to show that we're not quite done. Note that model 0 is the original model and model 1 is the model with the additional regressors, which will have a smaller residual (so this F can never be negative).

To get from approximately equal to an equals sign, we need to make it a bit like an elasticity – what is the percent change in the number of variables in the model? Suppose that we have N observations and that the original model has K variables, to which we're considering adding Q more observations. Then the original model has $(N - K - 1)$ degrees of freedom [that "1" is for the constant term] while the new model has $(N - K - Q - 1)$ degrees of freedom, so the difference is Q. So the percent change in degrees of freedom is $\frac{Q}{N - K - Q - 1}$. Then the full formula for the F test is

$$ F = \left(\frac{SSR_0 - SSR_1}{SSR_1} \right) \Bigg/ \left(\frac{Q}{N - K - Q - 1} \right). $$

Which is, admittedly, fugly, but perhaps similar enough to elasticity formulas to seem vaguely reassuring. But we know its distribution, it's F with (Q, N-K-Q-1) degrees of freedom – the F-distribution has 2 sets of degrees of freedom. Calculate that F, then use R to find pf(F, df1 =Q, df2 = (N-K-Q-1)) (or Excel to calculate FDIST(F,Q,N-K-Q-1)), to find a p-value for the test. If the p-value is less than 5%, reject the null hypothesis.

Usually you will have the computer spit out the results for you. In R, *anova (model1, model2)* or else *linearHypothesis ()* as we did before.

Don't be a dummy about Dummy Variables
It's important to think about the implicit restrictions imposed by the dummy specification – e.g. just putting in a dummy for "high school diploma or above" implicitly assumes that there are two groups, each relatively homogenous. So a regression of wage on just a dummy for high-school diploma assumes that there are two groups: those with a diploma and those without (many of whom have more than a high school degree) – and that each of these groups is relatively homogenous. In many cases the data might be too coarse to estimate fine distinctions: some datasets distinguish between people with a high school diploma and those with a GED while other data lump together those categories. (Many New Yorkers would distinguish which high school!) Every model makes certain assumptions but you want to consider them.
It might be wise to pack the education dummies into a factor and use that factor in R rather than playing around choosing to put in some but not all. This also takes care of automatically dropping one of the dummies (to use it as comparison). Consider these examples (which one is wrong?):

```
model1wrong <- lm(WSAL_VAL ~ educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv,  data = dat_use)
summary(model1wrong)
model2wrong <- lm(WSAL_VAL ~ educ_nohs + educ_hs + educ_smcoll + educ_as +
educ_bach + educ_adv,  data = dat_use)
summary(model2wrong)
model3wrong <- lm(WSAL_VAL ~ educ_hs + educ_bach,  data = dat_use)
summary(model3wrong)
```

In general it is better to use underlying continuous variables if you have them (e.g. for sports, net points scored rather than win/loss) – this is the basic intuition that there is no need to throw out information. On the other hand this imposes assumptions about linearity which might be inappropriate. For example,

```
model_continuousAge <- lm(WSAL_VAL ~ Age,  data = dat_use)
summary(model_continuousAge)
Age_factr <- cut(dat_use$Age,breaks=25:55)
model_factrAge <-lm(WSAL_VAL ~ Age_factr,  data = dat_use)
summary(model_factrAge)
plot(coef(model_factrAge))
```

Multiple Dummy Variables

Multiple dummy variables, $D_{1,i}$, $D_{2,i}$, ...,$D_{J,i}$, operate on the same basic principle. Of course we can then have many further interactions! Suppose we have dummies for education and immigrant status. The coefficient on education would tell us how the typical person (whether immigrant or native) fares, while the coefficient on immigrant would tell us how the typical immigrant (whatever her education) fares. An interaction of "more than Bachelor's degree" with "Immigrant" would tell how the typical highly-educated immigrant would do beyond how the "typical immigrant" and "typical highly-educated" person would do (which might be different, for both ends of the education scale).

Many, Many Dummy Variables

Often it is sensible to use lots of dummy variables. For example regressions to explain people's wages might use dummy variables for the industry in which a person works. Regressions about financial data such as stock prices might include dummies for the days of the week and months of the year.

Dummies for industries are often denoted with labels like "two-digit" or "three-digit" or similar jargon. To understand this, you need to understand how the government classifies industries. A specific industry might get a 4-digit code where each digit makes a further more detailed classification. The first digit refers to the broad section of the economy, as goods pass from the first producers (farmers and miners, first digit zero) to manufacturers (1 in the first digit for non-durable manufacturers such as meat processing, 2 for durable manufacturing, 3 for higher-tech goods) to transportation, communications and utilities (4), to wholesale trade (5) then retail (6). The 7's begin with FIRE (Finance, Insurance, and Real Estate) then services in the later 7 and early 8 digits while the 9 is for governments. The second and third digits give more detail: e.g. 377 is for sawmills, 378 for plywood and engineered wood, 379 for prefabricated wood homes. Some data sets might give you 5-digit or even 6-digit information. These classifications date back to the 1930s and 1940s so some parts show their age: the ever-increasing number of computer parts go where plain "office supplies" used to be.

The CPS data distinguishes between "major industries" with 16 categories and "detailed industry" with about 50.

Creating 50 dummy variables could be tiresome so that's where R's "$factor$" data type comes in handy. Just add in a factor into your OLS model and let R take care of the rest. So toss in A_DTIND and A_DTOCC. So add these lines and fire away,

```
det_ind <- as.factor(A_DTIND)
det_occ <- as.factor(A_DTOCC)
```

In other models such as predictions of sales, the specification might include a time trend (as discussed earlier) plus dummy variables for days of the week or months of the year, to represent the typical sales for, say, "a Monday in June".

Why are we doing all of this? Because I want you to realize all of the choices that go into creating a regression or doing just about anything with data. There are a host of choices available to you. Some choices are rather conventional (for example, the education breakdown I used above) but you need to know the field in order to know what assumptions are common. Sometimes these commonplace assumptions conceal important information. You want to do enough experimentation to understand which of your choices are crucial to your results. Then you can begin to understand how people might analyze the exact same data but come to varying conclusions. If your results contradict someone else's, then you have to figure out what are the important assumptions that create the difference.

If for example we wanted to have a dummy variable, D, interacted with a continuous variable, X, in a regression $y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3(D * X) + e$, we could write this formula in a number of different ways in R. Assuming D is already a factor (else might need to use `as.factor(D)`), we could write this as: `lm(y ~ X + D + X:D)` or, even more compactly, `lm(y ~ X*D)`. This allows each unit of D to have a different intercept as well as a different slope as X changes.

Panel Data

A panel of data contains repeated observations of a single economic unit over time. This might be a survey like the CPS where the same person is surveyed each month to investigate changes in their labor market status. There are medical panels that have given annual exams to the same people for decades. Publicly-traded firms that file their annual reports can provide a panel of data: revenue and sales for many years at many different firms. Sometimes data covers larger blocks such as states in the US or, if we're looking at macroeconomic development, even countries over time.

Other data sets are just cross-sectional, like the March CPS that we've used. If we put together a series of cross-sectional samples that don't follow the same people (so we use the March 2012, 2011, and 2010 CPS samples) then we have a pooled sample. A long stream of data on a single unit is a time series (for example US Industrial Production or the daily returns on a single stock).

In panel data we want to distinguish time from unit effects. Suppose that you are analyzing sales data for a large company's many stores. You want to figure out which stores are well-managed. You know that there are macro trends: some years are good and some are rough, so you don't want to indiscriminately reward everybody in good years (when they just got lucky) and punish them in bad years (when they got unlucky). There are also location effects: a store with a good location will get more traffic and sell more, regardless. So you might consider subtracting the average sales of a particular location away from current sales, to look at deviations from its usual. After doing this for all of the stores, you could subtract off the average deviation at a particular time, too, to account for year effects (if everyone outperforms their usual sales by 10% then it might just indicate a good economy). You would be left with a store's "unusual" sales – better or worse than what would have been predicted for a given store location in that given year.

A regression takes this even further to use all of our usual "prediction" variables in the list of X, and combine these with time and unit fixed effects.

Now the notation begins. Let the t-subscript index time; let j index the unit. So any observations of y and x must be at a particular date and unit; we have $y_{t,j}$ and then the k x-variables are each $x_{t,j}^{k}$ (the superscript for which of the x-variables). So the regression equation is

$$y_{t,j} = \alpha_j + \gamma_t + \beta_1 x_{t,j}^1 + \beta_2 x_{t,j}^2 + \ldots + \beta_{K-1} x_{t,j}^{K-1} + \beta_K x_{t,j}^K + e_{t,j},$$

where α_j (alpha) is the fixed effect for each unit j, γ_t (gamma) is the time effect, and then the error is unique to each unit at each time.

This is actually easy to implement, even though the notation might look formidable. Just create a dummy variable for each time period and another dummy for each unit and put the whole slew of dummies into the regression.

So, to take a tiny example, suppose you have 8 store locations over 10 years, 1999-2008. You have data on sales (Y) and advertising spending (X) and want to look at the relationship between this simple X and Y. So the data look like this:

$X_{1999,1}$	$X_{1999,2}$	$X_{1999,3}$	$X_{1999,4}$	$X_{1999,5}$	$X_{1999,6}$	$X_{1999,7}$	$X_{1999,8}$
$X_{2000,1}$	$X_{2000,2}$	$X_{2000,3}$	$X_{2000,4}$	$X_{2000,5}$	$X_{2000,6}$	$X_{2000,7}$	$X_{2000,8}$
$X_{2001,1}$	$X_{2001,2}$	$X_{2001,3}$	$X_{2001,4}$	$X_{2001,5}$	$X_{2001,6}$	$X_{2001,7}$	$X_{2001,8}$
$X_{2002,1}$	$X_{2002,2}$	$X_{2002,3}$	$X_{2002,4}$	$X_{2002,5}$	$X_{2002,6}$	$X_{2002,7}$	$X_{2002,8}$
$X_{2003,1}$	$X_{2003,2}$	$X_{2003,3}$	$X_{2003,4}$	$X_{2003,5}$	$X_{2003,6}$	$X_{2003,7}$	$X_{2003,8}$
$X_{2004,1}$	$X_{2004,2}$	$X_{2004,3}$	$X_{2004,4}$	$X_{2004,5}$	$X_{2004,6}$	$X_{2004,7}$	$X_{2004,8}$
$X_{2005,1}$	$X_{2005,2}$	$X_{2005,3}$	$X_{2005,4}$	$X_{2005,5}$	$X_{2005,6}$	$X_{2005,7}$	$X_{2005,8}$
$X_{2006,1}$	$X_{2006,2}$	$X_{2006,3}$	$X_{2006,4}$	$X_{2006,5}$	$X_{2006,6}$	$X_{2006,7}$	$X_{2006,8}$
$X_{2007,1}$	$X_{2007,2}$	$X_{2007,3}$	$X_{2007,4}$	$X_{2007,5}$	$X_{2007,6}$	$X_{2007,7}$	$X_{2007,8}$
$X_{2008,1}$	$X_{2008,2}$	$X_{2008,3}$	$X_{2008,4}$	$X_{2008,5}$	$X_{2008,6}$	$X_{2008,7}$	$X_{2008,8}$

and similarly for the Y-variables. To do the regression, create 9 time dummy variables: D2000, D2001, D2002, D2003, D2004, D2005, D2006, D2007, and D2008. Then create 7 unit dummies, D2, D3, D4, D5, D6, D7, and D8. Then regress the Y on X and these 16 dummy variables.

Then the interpretation of the coefficient on the X variable is the amount by which an increase in X, above its usual value for that unit and above the usual amount for a given year, would increase Y.

One drawback of this type of estimation is that it is not very useful for forecasting, either to try to figure out the sales at some new location or what will be sales overall next year – since we don't know either the new location's fixed effect (the coefficient on D9 or its alpha) or we don't know next year's dummy coefficient (on D2009 or its gamma).

We also cannot put in a variable that varies only on one dimension – for example, we can't add any other information about store location that doesn't vary over time, like its distance from the other stores or other location information. All of that variation is swept up in the firm-level fixed effect. Similarly we can't include macro data that doesn't vary across firm locations like US GDP since all of that variation is collected into the time dummies.

You can get much fancier; there is a whole econometric literature on panel data estimation methods. But simple fixed effects, put into the same OLS regression that we've become accustomed to, can actually get you far.

Multi-Level Modeling

After Fixed Effects, we can generalize to Multi-Level Modeling (much of my explanation is based on the excellent book, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, by Andrew Gelman & Jennifer Hill). From the wage regressions based on CPS data that we were using, we can consider adding information about the person's occupation (the data gives a rough grouping of people into about 20 occupations). You've probably done a version of this regression in your head, if you've ever read someone's job title and tried to figure out how much she makes.

There are a few ways to use the occupation data. One way is to ignore it, to not use it – which is what we were doing when we left it out of the regression. Everyone started from the same value. Gelman & Hill call this the "pooling" estimator since it pools everyone together. Another way would be to put in fixed effects for each occupation, letting each vary as needed – every occupation has a different intercept term, starting from a different value. This is "no-pooling." This puts no constraints at all on what the intercepts might be – some high, some low, some way far afield. A multilevel model imposes a model on how those intercepts vary: usually that they have a normal distribution with a central mean and variance. The math to define the estimator gets a bit more complicated, but we let the computer worry about that. But it's basically a weighted average of the "pooled" and "no-pooled" estimates, where the number of people reporting being in that particular group give the weights. So groups with a lot of members get nearly that "no-pooled" estimate, while a group with few members would be estimated to be like the larger group.

So in this example, the pooling case has wages of person i in industry j explained as $w_{i,j} = \alpha + \beta X_{i,j} + e_{i,j}$ (where the X includes all the rest of the variables, lumped together). The no-pooling case has $w_{i,j} = \alpha_j + \beta X_{i,j} + e_{i,j}$ so the intercept varies by industry, j. The multilevel case has $w_{i,j} = \alpha_0 + \alpha_{[j]} + \beta X_{i,j} + e_{i,j}$ but $\alpha_{[j]} \sim N(0, \sigma_\alpha)$.

With just a single level (like Occupation) this doesn't seem like a big thing, but if we want to define a lot of levels (Occupation, Industry, State or even City) then this gets more important. Instead of estimating a separate parameter for each level, we can estimate just overall parameters – and levels with only a small number of observations will be partially pooled.

Once we decide we want to do such a thing, the remaining question is, "how?" With R it's easy, just `lmer()` instead of `lm()`.

```
modelmm1 <- lmer(WSAL_VAL ~ as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)), dat_use)
summary(modelmm1)

modelmm2 <- lmer(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian + race_oth
            + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach + educ_adv
            + married + divwidsep + union_m + veteran + immigrant + immig2gen
          + as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)), dat_use)
summary(modelmm2)
```

In these cases we can compute the Intra-Class Correlation (ICC) which is the ratio of the variance in the groups (σ_α) to the total variance, so $\frac{\sigma_\alpha}{\sigma_\alpha + \sigma_\epsilon}$. Kind of like R², this goes from zero to one and is graded on a curve. It tells how important the within-group variation is, relative to the total variation.

Of course the next step would be to expand these coefficient estimates to be for slope as well as intercept – something like $w_{i,j} = \alpha_0 + \alpha_{[j]} + (\beta_0 + \beta_{[j]})X_{i,j} + e_{i,j}$. Multilevel modeling is a growing trend within econometrics.

```
# cps_1.R
# looking at CPS 2013 data
# uses file from dataferret download of CPS March 2013 supplement,
downloaded June 12 2014
# accompanying lecture notes for KFoster class ECO B2000 in fall 2014 at
CCNY

rm(list = ls(all = TRUE))
setwd("C:\\Users\\Kevin\\Documents\\CCNY\\data for classes\\CPS_Mar2013")
load("cps_mar2013.RData")

attach(dat_CPSMar2013)
# use prime-age,fulltime, yearround workers
use_varb <- (Age >= 25) & (Age <= 55) & work_fullt & work_50wks
dat_use <- subset(dat_CPSMar2013,use_varb) # 47,550 out of 202,634 obs

detach(dat_CPSMar2013)

attach(dat_use) # just prime-age,fulltime, yearround workers

# always a good idea to get basic stats of all of the variables in your
regression to see if they make sense
summary(WSAL_VAL)
summary(Age)
summary(female)
summary(AfAm)
summary(Asian)
summary(Amindian)
summary(race_oth)
summary(Hispanic)
summary(educ_hs)
summary(educ_smcoll)
summary(educ_as)
summary(educ_bach)
summary(educ_adv)
summary(married)
summary(divwidsep)
summary(union_m)
summary(veteran)
summary(immigrant)
summary(immig2gen)

model1 <- lm(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian + race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen)

summary(model1)
coeftest(model1)
#sometimes log form is preferred
# dat_noZeroWage <- subset(dat_use,(WSAL_VAL > 0))
# model1a <- lm(log(WSAL_VAL) ~ Age + female + AfAm + Asian + Amindian +
race_oth
#              + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
#              + married + divwidsep + union_m + veteran + immigrant +
immig2gen, data = dat_noZeroWage)
# detach(dat_use)
# attach(dat_noZeroWage)
# log(mean(WSAL_VAL))
# mean(log(WSAL_VAL))
# detach(dat_noZeroWage)
```

```r
# attach(dat_use)
# ^^ yes there are more elegant ways to do that, avoiding attach/detach -
find them!

# for heteroskedasticity consistent errors
require(sandwich)
require(lmtest)

coeftest(model1,vcovHC)

# jam nonlinear into linear regression
model2 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + AfAm + Asian + Amindian +
race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen)

model3 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + I(female*Age) +
I(female*(Age^2)) + AfAm + Asian + Amindian + race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen)
# could do this with "update" function instead
summary(model2)
summary(model3)
# the ANOVA function is flexible - can compare nested models
anova(model1,model2,model3)

# Applied Econometrics in R suggests also spline and kernel estimators, we
might get to that later

# subset in order to plot...
NNobs <- length(WSAL_VAL)
set.seed(12345) # just so you can replicate and get same "random" choices
graph_obs <- (runif(NNobs) < 0.1) # so something like 4000 obs
dat_graph <-subset(dat_use,graph_obs)

plot(WSAL_VAL ~ jitter(Age, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5,
alpha = 0.02), data = dat_graph)
# ^^ that looks like crap since Wages are soooooooo skew!  So try to find
ylim = c(0, ??)
plot(WSAL_VAL ~ jitter(Age, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5,
alpha = 0.02), ylim = c(0,150000), data = dat_graph)

# to plot the predicted values might want to do something like,
lines(fitted.values(model2) ~ Age)
# but that will plot ALLLLL the values, which is 4500 too many and looks
awful
# so back to this,
to_be_predicted2 <- data.frame(Age = 25:55, female = 1, AfAm = 0, Asian = 0,
Amindian = 1, race_oth = 1,
                               Hispanic = 1, educ_hs = 0, educ_smcoll = 0,
educ_as = 0, educ_bach = 1, educ_adv = 0,
                               married = 0, divwidsep =0, union_m = 0,
veteran = 0, immigrant = 0, immig2gen = 1)
to_be_predicted2$yhat <- predict(model2, newdata = to_be_predicted2)

lines(yhat ~ Age, data = to_be_predicted2)

# now compare model3
```

```
to_be_predicted3m <- data.frame(Age = 25:55, female = 0, AfAm = 0, Asian =
0, Amindian = 1, race_oth = 1,
                                Hispanic = 1, educ_hs = 0, educ_smcoll = 0,
educ_as = 0, educ_bach = 1, educ_adv = 0,
                                married = 0, divwidsep =0, union_m = 0,
veteran = 0, immigrant = 0, immig2gen = 1)
to_be_predicted3m$yhat <- predict(model3, newdata = to_be_predicted3m)

to_be_predicted3f <- data.frame(Age = 25:55, female = 1, AfAm = 0, Asian =
0, Amindian = 1, race_oth = 1,
                                Hispanic = 1, educ_hs = 0, educ_smcoll = 0,
educ_as = 0, educ_bach = 1, educ_adv = 0,
                                married = 0, divwidsep =0, union_m = 0,
veteran = 0, immigrant = 0, immig2gen = 1)
to_be_predicted3f$yhat <- predict(model3, newdata = to_be_predicted3f)

plot(WSAL_VAL ~ jitter(Age, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5,
alpha = 0.02), ylim = c(0,150000), xlab = "Age", data = dat_graph)
lines(yhat ~ Age, data = to_be_predicted3f)
lines(yhat ~ Age, data = to_be_predicted3m, lty = 2)
legend("topleft", c("male", "female"), lty = c(2,1), bty = "n")

det_ind <- as.factor(A_DTIND)
det_occ <- as.factor(A_DTOCC)

model4 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + AfAm + Asian + Amindian +
race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen
             + det_ind + det_occ)
summary(model4)

# and always remember this part...
detach(dat_use)
```

```
# cps_2.R
# looking at CPS 2013 data
# uses file from dataferret download of CPS March 2013 supplement,
downloaded June 12 2014
# accompanying lecture notes for KFoster class ECO B2000 in fall 2014 at
CCNY

rm(list = ls(all = TRUE))
setwd("C:\\Users\\Kevin\\Documents\\CCNY\\data for classes\\CPS_Mar2013")
load("cps_mar2013.RData")

attach(dat_CPSMar2013)
# use prime-age,fulltime, yearround workers
use_varb <- (Age >= 25) & (Age <= 55) & work_fullt & work_50wks
dat_use <- subset(dat_CPSMar2013,use_varb) # 47,550 out of 202,634 obs

detach(dat_CPSMar2013)

# create a single index variable (factor) from education dummies
# educ_indx <- as.factor(educ_nohs + 2*educ_hs + 3*educ_smcoll + 4*educ_as +
5*educ_bach + 6*educ_adv)
# levels(educ_indx)[1] <- "No HS"
# levels(educ_indx)[2] <- "HS"
# levels(educ_indx)[3] <- "Some Coll"
# levels(educ_indx)[4] <- "AS"
```

```r
# levels(educ_indx)[5] <- "Bach"
# levels(educ_indx)[6] <- "Adv Deg"
# levels(educ_indx)

attach(dat_use) # just prime-age,fulltime, yearround workers
# will look at some info by industry so look how wage varies by ind:
by(WSAL_VAL, A_DTOCC, summary)
plot(as.factor(female) ~ A_DTOCC)

detach(dat_use)
# A_DTOCC values:
# 1 'Management occupations'
# 2 'Business and financial operations occupations'
# 3 'Computer and mathematical science occupations'
# 4 'Architecture and engineering occupations'
# 5 'Life, physical, and social service occupations'
# 6 'Community and social service occupations'
# 7 'Legal occupations'
# 8 'Education, training, and library occupations'
# 9 'Arts, design, entertainment, sports, and media occupations'
# 10 'Healthcare practitioner and technical occupations'
# 11 'Healthcare support occupations'
# 12 'Protective service occupations'
# 13 'Food preparation and serving related occupations'
# 14 'Building and grounds cleaning and maintenance occupations'
# 15 'Personal care and service occupations'
# 16 'Sales and related occupations'
# 17 'Office and administrative support occupations'
# 18 'Farming, fishing, and forestry occupations'
# 19 'Construction and extraction occupations'
# 20 'Installation, maintenance, and repair occupations'
# 21 'Production occupations'
# 22 'Transportation and material moving occupations'
# 23 'Armed Forces'




# for heteroskedasticity consistent errors
require(sandwich)
require(lmtest)


model1 <- lm(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian + race_oth
             + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
             + married + divwidsep + union_m + veteran + immigrant +
immig2gen,  data = dat_use)
summary(model1)
coeftest(model1,vcovHC)

# can do it wrong...
model1wrong <- lm(WSAL_VAL ~ educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv,  data = dat_use)
summary(model1wrong)
model2wrong <- lm(WSAL_VAL ~ educ_nohs + educ_hs + educ_smcoll + educ_as +
educ_bach + educ_adv,  data = dat_use)
summary(model2wrong)
model3wrong <- lm(WSAL_VAL ~ educ_hs + educ_bach,  data = dat_use)
summary(model3wrong)
# model1 leaves out varbs;
# model2 creates perfect multicollinearity with too many dummies;
# model3 has too few dummies
```

```
# example with Age
model_continuousAge <- lm(WSAL_VAL ~ Age,  data = dat_use)
summary(model_continuousAge)
Age_factr <- cut(dat_use$Age,breaks=25:55)
model_factrAge <-lm(WSAL_VAL ~ Age_factr,  data = dat_use)
summary(model_factrAge)
plot(coef(model_factrAge))

model2 <- lm(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian + race_oth
            + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
            + married + divwidsep + union_m + veteran + immigrant +
immig2gen
            + as.factor(A_DTOCC), data = dat_use)
summary(model2)
coeftest(model2,vcovHC)

require(lme4)
# next use multilevel based on industry A_DTOCC
modelmm1 <- lmer(WSAL_VAL ~ as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)),
dat_use)
summary(modelmm1)

modelmm2 <- lmer(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian +
race_oth
              + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
educ_adv
              + married + divwidsep + union_m + veteran + immigrant +
immig2gen
            + as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)), dat_use)
summary(modelmm2)
```

Instrumental Variables

- Endogenous vs. Exogenous variables

  o Exogenous variables are generated from "exo" outside of the model; endogenous are generated from "endo" within the model. Of course this neat binary distinction rarely is matched by the world; some variables are more endogenous than others

- Data can only demonstrate correlations – we need theory to get to causation. "Correlation does not imply causation." Roosters don't make the sun rise. Although Granger Causation from the logical inverse: not-correlate implies not-cause. If knowledge of variable X does not help predict Y, then X does not cause Y.

- In any regression, the variables on the right-hand side should be exogenous while the left-hand side should be endogenous, so X causes Y, $X \rightarrow Y$. But we should always ask if it might be plausible for Y to cause X, $Y \rightarrow X$, or for both X and Y to be caused by some external factor. If we have a circular chain of causation (so $X \rightarrow Y$ and $Y \rightarrow X$) then the OLS estimates are meaningless for describing causation. (So often need to watch dates – if the X variables are date (t-1) while Y is date t, then the causation is clearer than if all are dated t.) Example: oil prices and economic growth – high oil prices can choke off growth, but lower growth means less demand so lower oil prices.

- **NEVER** regress Price on a Quantity or vice versa!

Why never regress Price on Quantity? Wouldn't this give us a demand curve? Or, would it give us a supply curve? Why would we expect to see one and not the other?

In actuality, we don't observe either supply curves or demand curves. We only observe the values of price and quantity where the two intersect.

If both vary randomly then we will not observe a supply or a demand curve. We will just observe a cloud of random points.

For example, theory says we see this:
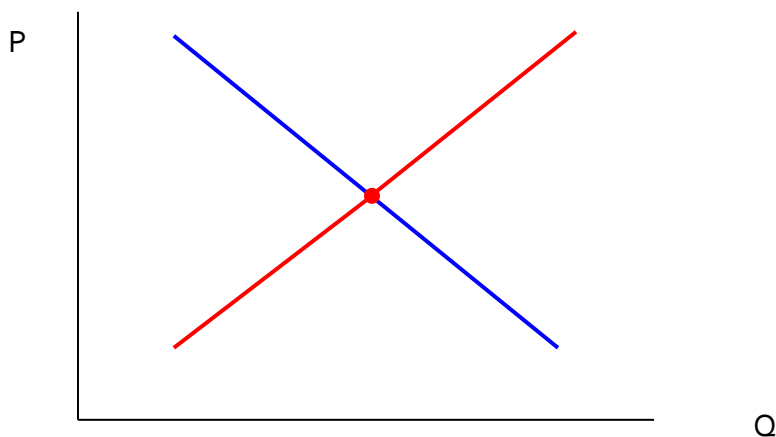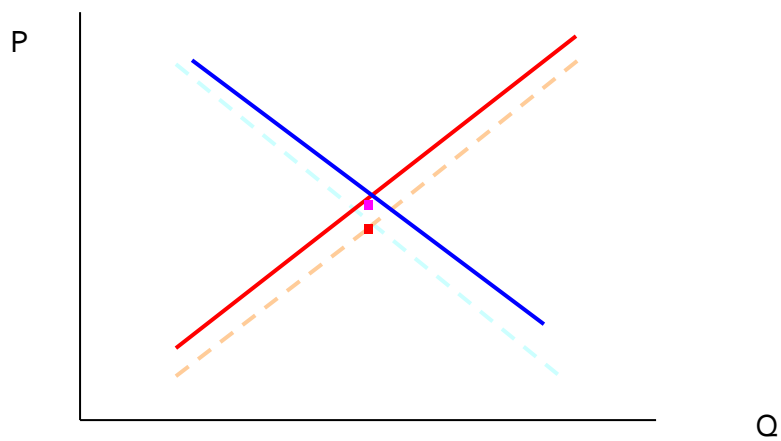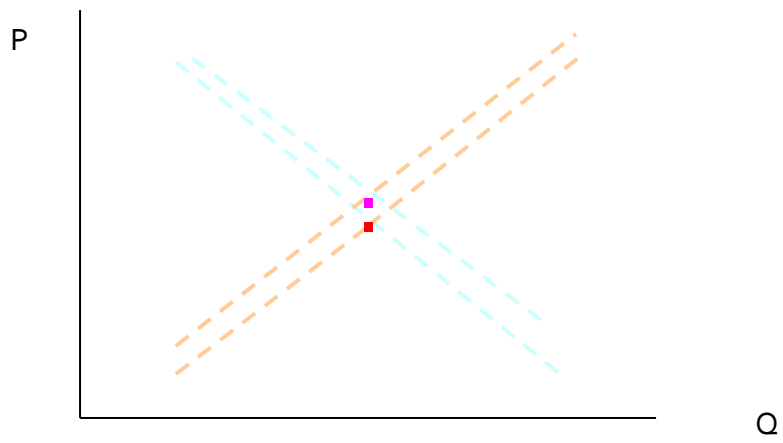

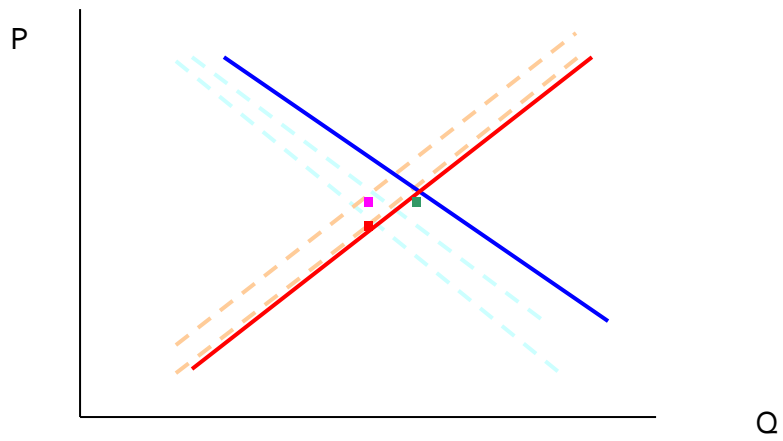But in the world, we assume the dotted-lines and only actually observe the one intersection, the dot:
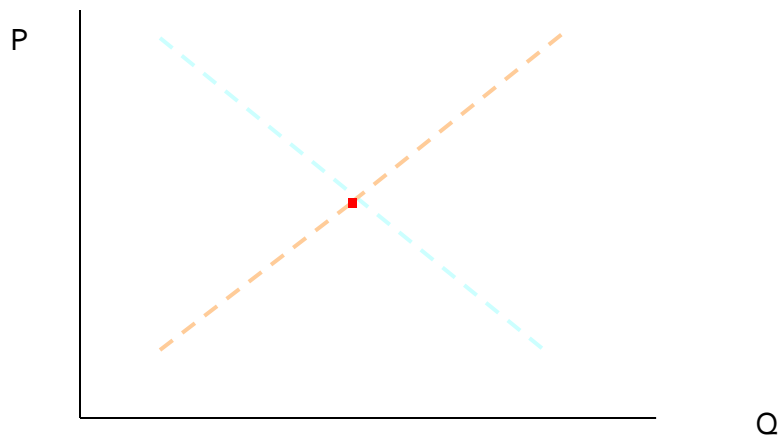
In the next time period, supply and demand shift randomly (well, for some reason we don't know, so it's random to us) by a bit, so theory tells us that we now have:
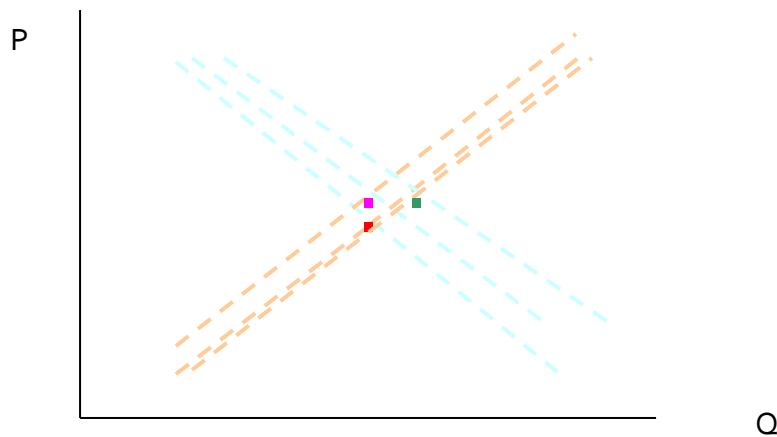

But again we actually now see just two points,

In the third period,

but again, in actuality just three points:

P

Q

So if we tried to draw a regression line on these three points, we might convince ourselves that we had a supply curve. But do we, really? You should be able to re-draw the lines to show that we could have a down-sloping line, or a line with just about any other slope.

So even if we could wait until the end of time and see an infinite number of such points, we'd still never know anything about supply and demand. This is the problem of endogeneity. The regression is **not identified** – we could get more and more information but still never learn anything.

We could show this in an Excel sheet or little R program, too, which will allow a few more repetitions.

Recall that we can write a demand curve as $P_d = A - BQ_d$ and a supply curve as $P_s = C + DQ_s$, where generally A, B, C, and D are all positive real numbers. In equilibrium $P_d = P_s$ and $Q_d = Q_s$. For simplicity assume that A=10, C=0, and B=D=1. Without any randomness this would be a boring equation; solve to find $10 - Q = Q$ and Q*=5, P*=5. (You did this in Econ 101 when you were a baby!) If there were a bit of randomness then we could write $P_d = A - BQ_d + \varepsilon_d$ and $P_s = C + DQ_s + \varepsilon_s$. Now the equilibrium conditions tell that

$$10 - Q + \varepsilon_d = Q + \varepsilon_s \text{ and so } Q^* = \frac{10 + \varepsilon_d - \varepsilon_s}{2} = 5 + \frac{(\varepsilon_d - \varepsilon_s)}{2} \text{ and } P^* = 5 + \frac{\varepsilon_d - \varepsilon_s}{2} + \varepsilon_s = 5 + \frac{\varepsilon_d + \varepsilon_s}{2}.$$

Plug this into Excel, assuming the two errors are normally distributed with mean zero and standard deviation of one, and find that the scatter looks like this (with 30 observations):
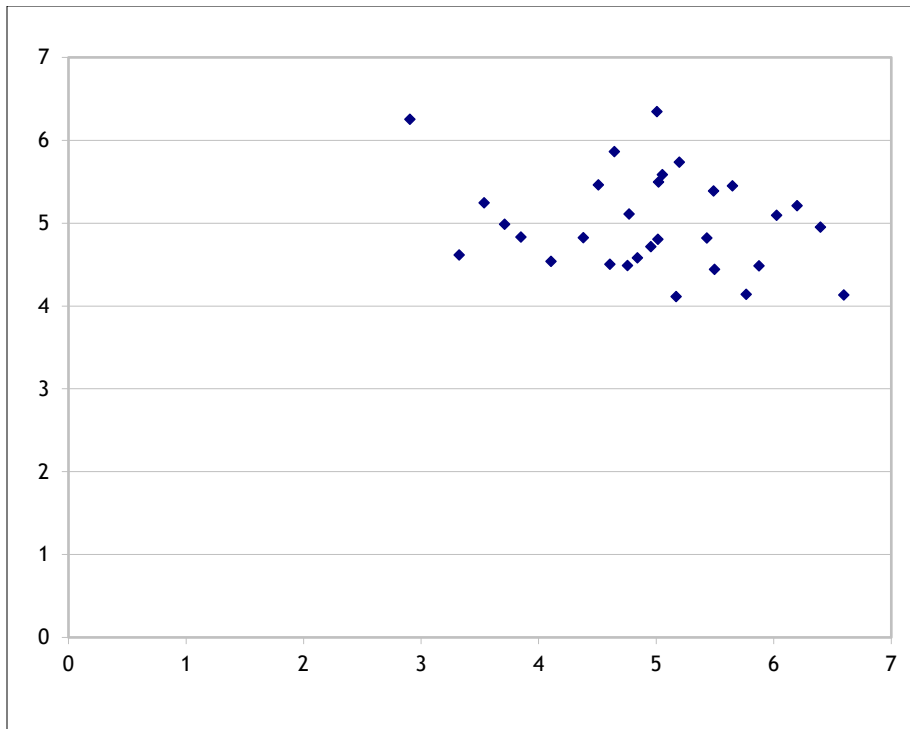
You can play with the spreadsheet, hit F9 to recalculate the errors, and see that there is no convergence, there is nothing to be learned.

On the other hand, what if the supply errors were smaller than the demand errors, for example by a factor of 4? Then we would see something like this:
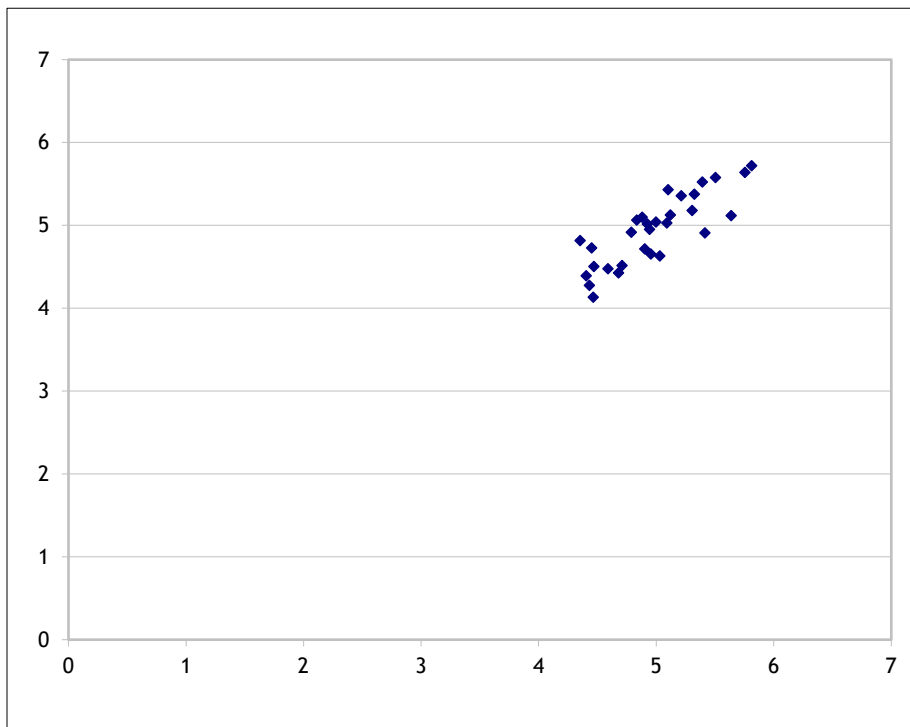
So now with the supply curve pinned down (relatively) we can begin to see it emerge as the demand curve varies.

But note the rather odd feature: variation in demand is what identifies the supply curve; variation in supply would likewise identify the demand curve. If some of that demand error were observed then that would help identify the supply curve, or vice versa. So sometimes in agricultural data we would use weather variation (that affects the supply of a commodity) to help identify demand.

If you want to get a bit crazier, experiment if the slope terms have errors as well (so $P_d = (A + \varepsilon_a) - (B + \varepsilon_b)Q_d$ and $P_s = (C + \varepsilon_c) + (D + \varepsilon_D)Q_s$).

Check Hal Varian http://www.pnas.org/content/113/27/7310.long

Instrumental Variables Regression

The basic idea of instrumental variables is that if we have some regression,

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

But X and Y are endogenous, then suppose we had some variable Z, which is uncorrelated with Y but still explains X, then we can make a supplementary regression,

$$X = \gamma_0 + \gamma_1 Z + u,$$

And get $\hat{X}$, the predicted values from that regression, then do the original regression as

$$Y = \beta_0 + \beta_1 \hat{X} + \varepsilon,.$$

- valid instrument, some Z_i for regression $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + u_i$
  - relevance: $corr(Z_i, X_i) \neq 0$ and
  - exogeneity: $corr(Z_i, u_i) = 0$
  - instrument explains X but NOT Y – can be excluded from list of variables explaining Y
- Two-Stage Least Squares (TSLS or 2SLS)
  - $X_i = \pi_0 + \pi_1 Z_i + v_i$, $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - get $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ and regress Y_i on $\hat{X}_i$
  - $\hat{\beta}_1 = \dfrac{s_{ZY}}{s_{ZX}}$
- General Case:
  - $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$
    $+ \hat{\beta}_{k+1} W_{1i} + \hat{\beta}_{k+2} W_{2i} + \ldots + \hat{\beta}_{k+r} W_{ri} + u_i$
  - X are endogenous regressors
  - W are exogenous regressors
  - $Z_{1i}, Z_{2i}, \ldots Z_{mi}$ are instruments
  - if m>k then "overidentified"; if m=k then just identified; if m<k then unidentified
  - still need:
    - $E(u_i | W_{1i}, W_{2i}, \ldots W_{ri}) = 0$
    - X, W, Z are all i.i.d. with fourth moments
    - W not perfectly collinear

- Instrument Relevance and Exogeneity
- Two-Stage Least Squares:
  - regress X on Z to get $\hat{X}$
  - then regress Y on W and $\hat{X}$
- Evaluating Instruments in the Real World
  - Weak instruments: check first-stage regression F-stat bigger than 10?
  - Examples:
    - cigarette tax to find effect of price
    - prison capacity in place of jail terms
    - random variation in births for class size
    - geography for heart attack treatment
    - number of immigrants 10 years ago for immigrant increase
    - Mariel boatlift, other policy shifts
    - deployment of police after 9/11 to estimate effects of police on crime
  - Bad examples of poor instruments:
    - weak instrument: month of birth on wage earnings
  - Many bad examples where instruments needed:
    - wage explained by schooling
    - health insurance explained by wage
    - wage explained by weight (discrimination against fat people?) vs wage explained by race/ethnicity (discrimination against minorities)
- Heckman 2-step for 2-part questions: first, "yes or no?"; next "how much?" Like 2SLS but first stage is a probit (we'll do that later)! Again need an exclusion restriction, some variable that explains the first step but not the second.

Instrumental Variables Regression in R

There was a paper in the journal *Economic Inquiry*, by Cesur & Kelly (2013), "Who Pays the Bar Tab? Beer Consumption and Economic Growth in the United States," which concluded that beer consumption was bad for economic growth. I got data from the Brewer's Almanac, provided online by the Beer Institute (beerinstitute.org) and the Bureau of Economic Analysis (bea.gov). This is not quite the same data that the paper used (less complete) but it gives a flavor (bad pun) of the results.

You can download the R data from the class webpage. Then run this regression,
```
regression1 <- lm(growth_rates ~ beer_pc + gdp_L + as.factor(st_fixedeff))
summary(regression1)
```

Where the growth rate of each state's GDP is a function of per-capita beer consumption, a lag of state GDP (reflecting the general idea that poorer states might grow faster), as well as state fixed effects (each state has its own intercept). This shows a positive and statistically significant coefficient on per-capita beer consumption. So beer is good for growth?!

As Homer Simpson put it, "To alcohol! The cause of – and solution to – all of life's problems." That circularity of causation makes the statistics more complicated.

Richer people have more money to buy everything including beer, so economic growth might cause beer consumption. One way out, suggested by the article authors, is to use an instrument for beer consumption – the tax on beer. This is a plausible instrument since it likely causes changes in beer consumption (higher price, lower consumption, y'know the demand curve) but it unlikely to be affected by economic growth. So estimate an instrumental variables equation,

```
iv_reg1 <- lm(beer_pc ~ beertax)
summary(iv_reg1)
```

And see that indeed there is a negative coefficient (hooray for demand curves!) although it is certainly a weak instrument (R^2 less than 1%). Use the predicted value of beer consumption per capita as an instrument in the regression in place of the endogenous variable,

```
pred_beer <- predict(iv_reg1)
iv_reg2 <- lm(growth_rates ~ pred_beer + gdp_L + as.factor(st_fixedeff))
summary(iv_reg2)
```

To note that now beer consumption seems to have negative effects on economic growth (only significant at 10% level; the article adds some other variables to get it significant[1]). I put some other variables in the dataset that you might play with – see if you can find the opposite result! (R code from a simple summary at http://www.r-bloggers.com/a-simple-instrumental-variables-problem/)

Finally note that you can use the AER package and ivreg() procedure for better results, since these estimated standard errors won't be quite right – but that's just fine-tuning.

Measuring Discrimination – Oaxaca Decompositions:

(much of this discussion is based on Chapter 10 of George Borjas' textbook on Labor Economics)

---

[1] Actually the published article uses the real level of beer tax, so they divide by the state-level CPI. IMHO that makes it a poor instrument since there is no reason to think that price level is exogenous with GDP - in fact there are centuries of economists back at least to Adam Smith saying that price level and incomes are closely related! But that's just my opinion, clearly other people believe otherwise.

The regressions that we've been using measured the returns to education, age, and other factors upon the wage. If we classify people into different groups, distinguished by race, ethnicity, gender, age, or other categories, we can measure the difference in wages earned. There are many explanations but we want to determine how much is due to discrimination and how much due to different characteristics (chosen or given).

Consider a simple model where we examine the native/immigrant wage gap, and so measure $\overline{w}_N$, the average wages that natives get, and $\overline{w}_M$, the average wages that immigrants get. The simple measure, $\overline{w}_N - \overline{w}_M$, of the wage gap, would not be adequate if natives and migrants differ in other ways, as well.

Consider the effect of age. Theory implies that people choose to migrate early in life, so we might expect to see age differences between the groups. And of course age influences the wage. If natives and immigrants had different average wages solely because of having different average ages, we would conclude very different reasons for this than if the two groups had identical ages but different wages.

For example, in a toy-sized 1000-observation subset of CPS March 2005 data, there are 406 natives and 77 immigrants workers with non-zero wages. The natives averaged wage/salary of $37,521 while the immigrants had $32,507. The average age of the natives was 39.5; the average age of the immigrants was 42.1. We want to know how much of the difference in wage can be explained by the difference in age.

Consider a simple model that posits different simple regressions for natives and immigrants:

$$w_N = \beta_{0,N} + \beta_{1,N} Age + \varepsilon$$

$$w_M = \delta_{0,M} + \delta_{1,M} Age + \varepsilon .$$

We know that average wages for natives depend on average age of natives, $\overline{Age}_N$:

$$\overline{w}_N = \beta_{0,N} + \beta_{1,N} \overline{Age}_N$$

and for immigrants as well, wages depend on immigrants' average age, $\overline{Age}_M$:

$$\overline{w}_M = \delta_{0,M} + \delta_{1,M} \overline{Age}_M .$$

The difference in average wages is:

$$\overline{w}_N - \overline{w}_M = \left(\beta_{0,N} + \beta_{1,N} \overline{Age}_N \right) - \left(\delta_{0,M} + \delta_{1,M} \overline{Age}_M \right)$$

but we can add and subtract the cross term, $\delta_{1,M} \overline{Age}_N$ to get:

$$\overline{w}_N - \overline{w}_M = \left(\beta_{0,N} - \delta_{0,M} \right) + \left(\beta_{1,N} - \delta_{1,M} \right) \overline{Age}_N + \delta_{1,M} \left(\overline{Age}_N - \overline{Age}_M \right) .$$

Each term can be interpreted in different ways. The first difference, $\left(\beta_{0,N} - \delta_{0,M}\right)$, is the difference in intercepts, the parallel shift of wages for all ages. The second, $\left(\beta_{1,N} - \delta_{1,M}\right)\overline{Age}_N$, is the difference in how the skills are rewarded: if everyone in the data were to have the same age, immigrants and natives would still have different wages due to these first two factors. The third is $\delta_{1,M}\left(\overline{Age}_N - \overline{Age}_M\right)$, which gives the difference in wage attributable only to differences in average age (even if those were rewarded equally). The first two are generally regarded as due to discrimination while the last is not.

The basic framework can be extended to other observable differences: in years of education, experience, or the host of other qualifications that affect people's wages and salaries.

From our discussions of regression models, we realize that the two equations above could be combined into a single framework. If we define an immigrant dummy variable as M_i, which is equal to one if individual i is an immigrant and zero if that person is native born, we can write a regression model as:

$$w_i = \beta_0 + \beta_1 Age_i + \beta_2 M_i + \beta_3 M_i Age_i + \varepsilon_i,$$

where wages for natives depend on only β_0 and β_1, while the immigrant coefficients are $\delta_{0,M} = \beta_0 + \beta_2$ and $\delta_{1,M} = \beta_1 + \beta_3$. We construct $\overline{w}_N = \hat{\beta}_0 + \hat{\beta}_1 \overline{Age}_N$ and $\overline{w}_M = \hat{\beta}_0 + \hat{\beta}_2 + \left(\hat{\beta}_1 + \hat{\beta}_3\right)\overline{Age}_M$ so the Oaxaca decomposition is now:

$$\overline{w}_N - \overline{w}_M = -\beta_2 - \beta_3 \overline{Age}_N + \left(\beta_1 + \beta_3\right)\left(\overline{Age}_N - \overline{Age}_M\right).$$

We note that unobserved differences in quality of skills can be measured as instead being due to discrimination. In our example, suppose that natives get a greater salary as they age due to the skills which they amass, but immigrants who have language difficulties learn new skills more slowly. In this case, older natives would earn more, increasing the returns to aging. This would be reflected as lower coefficients on age for immigrants than natives, and so evidence of discrimination. If we had information on English-language ability (SAT, TOEFL or GRE scores, maybe?), then the regression would show that a lack of those skills led to lower wages – no longer would it be measured as evidence of discrimination.

But this elides the question of how people gain the "skills" measured in the first place. If a degree from a foreign university gets less reward than a degree from an American university, is this entirely due to discrimination? What fraction of the wage differential arises from skill differences? In the US, African-American and Hispanic children tend to go to lower-quality schools (as measured by test scores or teacher qualifications). The lower subsequent wages might not be due to labor market discrimination (if firms rationally pay less for lower skills) but still be due to societal discrimination.

Consider the sort of dataset that we've been working with. Regressing Age, an Immigrant dummy, and an Age-Immigrant interaction on Wage provides the following coefficient estimates (for the same sub-sample as before):

$$w_i = 7437 + 762.62 Age_i + 20,663.29 M_i - 658.06 Age_i M_i + \varepsilon_i$$

where the immigrant dummy is actually positive (neither the immigrant dummy nor the immigrant-age interaction term are statistically significant, but I ignore that for now). With the average ages from above (natives 39.5 years old; immigrants 42.1), we calculate the gap in average predicted wages (natives are predicted to make an average wage of \$37,561; immigrants to make \$32,502) is \$5058.08. The two first terms in the Oaxaca decomposition, relating to unexplained factors such as "discrimination" $-\hat{\beta}_2 - \hat{\beta}_3 \overline{Age}_N$ account for \$5329.95, while the difference in age accounts for just -\$271.86 (a negative amount) – this means that the ages actually imply that natives and immigrants ought to be closer in wages so they are even farther apart. We might reasonably believe that much of this difference reflects omitted factors (and could list out the important omitted factors); this is intended merely as an exercise.

Adding these additional variables is easy; I show the case for two variables but the model can be extended to as many variables as are of interest. Next consider a more complicated model, where now wages depend on Age and Education, so the two regressions for natives and immigrants are:

$$w_N = \beta_{0,N} + \beta_{1,N} Age + \beta_{2,N} Educ + \varepsilon$$

$$w_M = \delta_{0,M} + \delta_{1,M} Age + \delta_{2,M} Educ + \varepsilon .$$

We know that average wages for natives depend on average age and education of natives, $\overline{Age}_N, \overline{Educ}_N$:

$$\overline{w}_N = \beta_{0,N} + \beta_{1,N} \overline{Age}_N + \beta_{2,N} \overline{Educ}$$

and for immigrants as well, wages depend on immigrants' average age, $\overline{Age}_M, \overline{Educ}_M$:

$$\overline{w}_M = \delta_{0,M} + \delta_{1,M} \overline{Age}_M + \delta_{2,M} \overline{Educ} .$$

The difference in average wages is:

$$\overline{w}_N - \overline{w}_M = \left(\beta_{0,N} + \beta_{1,N} \overline{Age}_N + \beta_{2,N} \overline{Educ}_N \right) - \left(\delta_{0,M} + \delta_{1,M} \overline{Age}_M + \delta_{2,M} \overline{Educ}_M \right)$$

but we can add and subtract the cross terms , $\delta_{1,M} \overline{Age}_N + \delta_{2,M} \overline{Age}_N$ to get:

$$\overline{w}_N - \overline{w}_M = \left(\beta_{0,N} - \delta_{0,M} \right) + \left(\beta_{1,N} - \delta_{1,M} \right) \overline{Age}_N + \delta_{1,M} \left(\overline{Age}_N - \overline{Age}_M \right) + \left(\beta_{2,N} - \delta_{2,M} \right) \overline{Educ}_N + \delta_{2,M} \left(\overline{Educ}_N - \overline{Educ}_M \right)$$
.

Again, the two terms showing the difference in average levels of external factors, $\left(\overline{Age}_N - \overline{Age}_M \right)$ and $\left(\overline{Educ}_N - \overline{Educ}_M \right)$, are "explained" by the model while the other terms showing the difference in the coefficients are "unexplained" and could be considered as evidence of discrimination.

Exercises:
1.   Do the above analysis on the current CPS data.
2.   If instead you used log wages, but still kept just age as the measured variable, is your answer substantially different than in the previous question? (Note that the answers are in different units, so you have to think about how to convert the two answers.)

3.   Consider other measures of skills, such as schooling and whatever other factors you consider important. How does this new regression change the Oaxaca decomposition?

4.   What is the maximum fraction of wage difference that you can find (with different independent variables and regression specifications), related to discrimination? The minimum?
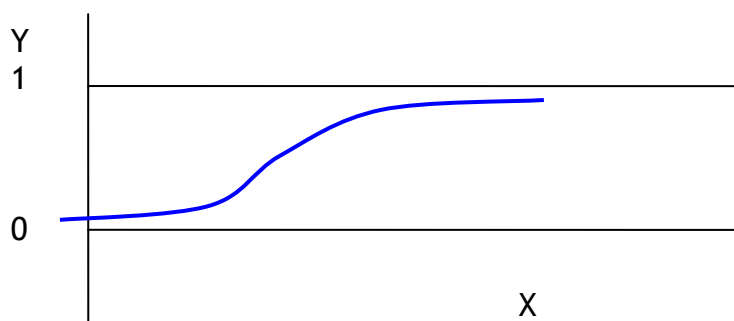
References:

Borjas, George (2003). *Labor Economics*.

Oaxaca, Ronald (1973). "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14(3).


Binary Dependent Variable Models

(Stock & Watson Chapter 9)

• Sometimes our dependent variable is continuous, like a measurement of a person's income; sometimes it is just a "yes" or "no" answer to a simple question. A "Yes/No" answer can be coded as just a 1 (for Yes) or a 0 (a zero for "no"). These zero/one variables are called dummy variables or **binary** variables. Sometimes the dependent variable can have a range of discrete values ("How many children do you have?" "Which train do you take to work?") – in this case we have a discrete variable. The binary and continuous variables can be seen as opposite ends of a spectrum.

• We want to explore models where our dependent variable takes on discrete values; we'll start with just binary variables. For example, we might want to ask what factors influence a person to go to college, to have health insurance, or to look for a job; to have a credit card or get a mortgage; what factors influence a firm to go bankrupt; etc.

• Linear Models such as OLS have some problems. These imply predicted values of Y that are greater than one or less than zero. They also have advantages! You should be able to do both
http://marcfbellemare.com/wordpress/8951

• Interpret our prediction of Y as being the probability that the Y variable will take a value of one. (Note: remember which value codes to one and which to zero – there is no necessary reason, for example, for us to code Y=1 if a person has health insurance; we could just as easily define Y=1 if a person is uninsured. The mathematics doesn't change but the interpretation does!)

• want to somehow "bend" the predicted Y-value so that the prediction of Y never goes above 1 or below zero, something like:

- Probit Model

  - $\Pr(Y=1|X)=\Phi(\beta_0+\beta_1X_1+\beta_2X_2)$ where $\Phi(\square)$ is the cdf of the standard normal

  - $\dfrac{\Delta\Pr}{\Delta X}$ is not constant

- Logit Model

  - $\Pr(Y=1|X)=F(\beta_0+\beta_1X_1+\beta_2X_2)$, where $F(z)=\dfrac{1}{1+e^{-z}}$

  - $\dfrac{\Delta\Pr}{\Delta X}$ is not constant

- differences (Excel sheet:compare_probit_logit.xls)

Clearly the differences are rather small; it is rare that we might have a serious theoretical justification for one specification rather than the other.
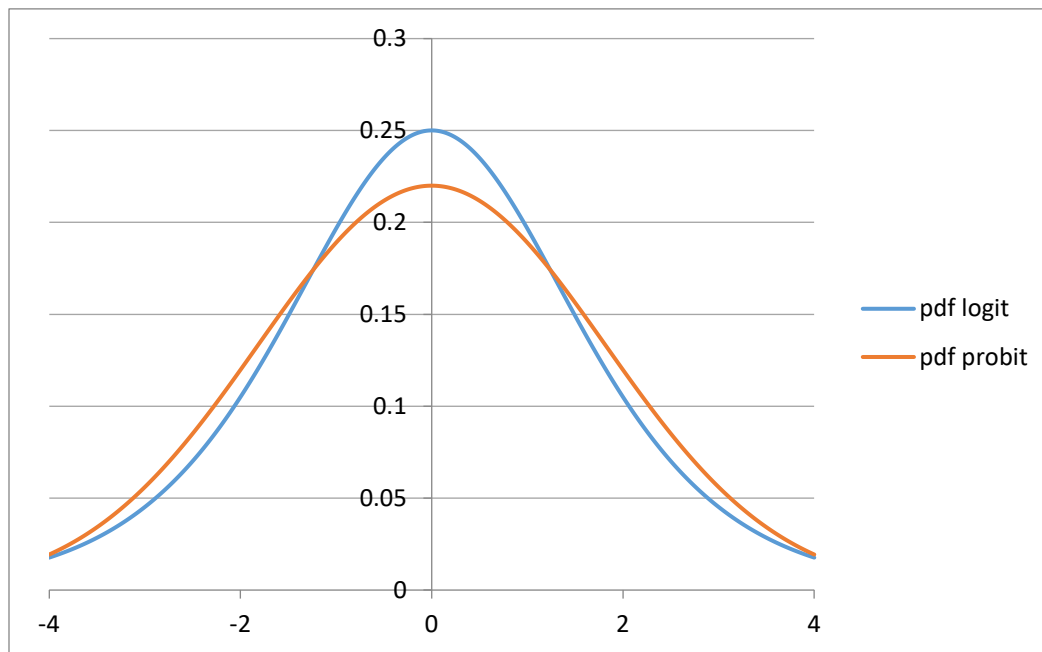
(Note that the logit function given above has standard error of $\dfrac{\pi}{\sqrt{3}}$ so in the plots I scaled the probit by this factor).

CDF
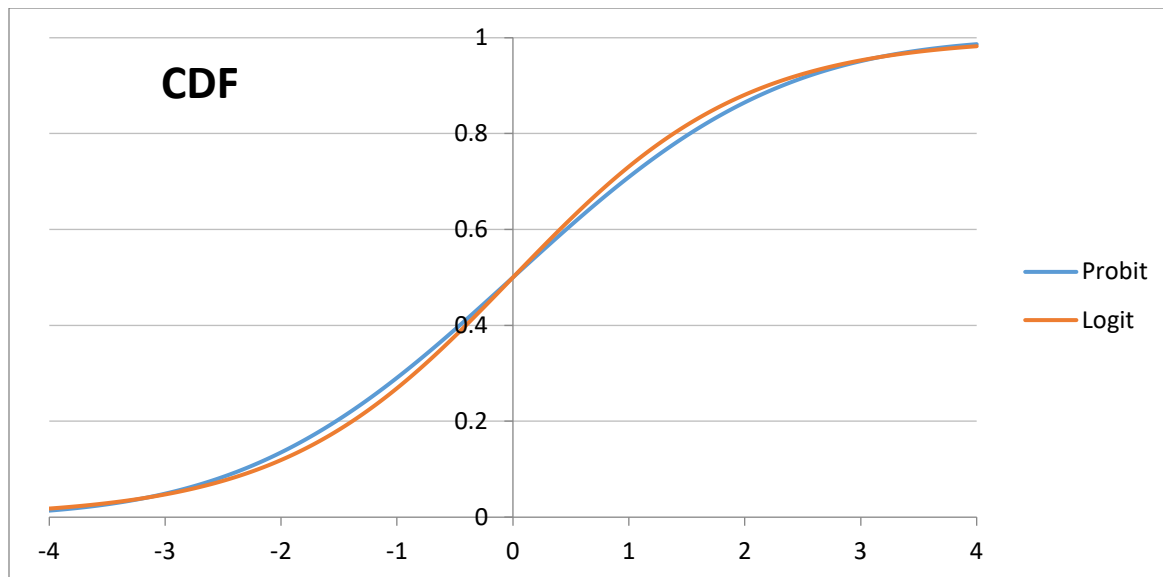
- Measures of Fit

  o no single measure is adequate; many have been proposed

  o What probability should be used as "hit"? If the model says there is a 90% chance of Y=1, and it truly is equal to one, then that is reasonable to count as a correct prediction. But many measures use 50% as the cutoff. Tradeoff of false positives versus false negatives – loss function might well be asymmetric.

	actually = 1	*actually = 0*
Predicted = 1	*Hooray!*	*sad*
Predicted = 0	*sad (maybe sadder?)*	*Hooray!*

Probit/Logit in R

For a logit estimation, just

```
regn_logit1 <- glm(Y ~ X1 + X2, family = binomial, data = data1)
```

for a probit estimation

```
regn_probit1 <- glm(Y ~ X1 + X2, family = binomial (link = 'probit'), data =
data1)
```

```
Example with CPS data

model_logit1 <- glm(health_ins ~ Age + I(Age^2) + female + AfAm +
      Asian + Amindian + race_oth + Hispanic + educ_hs + educ_smcoll +
      educ_as + educ_bach + educ_adv + married + divwidsep + union_m +
      veteran + immigrant + immig2gen, family = binomial, data =
      dat_use_hi)

summary(model_logit1)

regn_probit1 <- glm(health_ins ~ Age + female + AfAm + Asian +
      Amindian + race_oth + Hispanic + educ_hs + educ_smcoll + educ_as
      + educ_bach + educ_adv + married + divwidsep + union_m + veteran
```

```
            + immigrant + immig2gen, family = binomial (link = 'probit'),
            data = dat_use_hi)

summary(regn_probit1)
```

Then the estimation results from "`summary()`" should be familiar. The interpretation is also essentially unchanged: look at the individual t-statistics (formed by dividing coefficient estimates from standard errors) then get a p-value from that.

In addition to looking at effects of particular X-variables, we are interested in looking at predictive accuracy – but note that this is likely to vary depending on your project so the results I'm going to show here are particular to this analysis. You would have to carefully take a look at your own model predictions. Also would want to check different sub-groups – is predictive accuracy substantially better or worse for particular groups? That might be a signal that the simple dummies are not adequately capturing the variation.

```
summary(model_logit1$fitted)
summary(dat_use_hi$health_ins)
pred_model_logit1 <- (model_logit1$fitted > 0.5)
table(pred_model_logit1, dat_use_hi$health_ins)
frac_correct_l1a <- mean(as.numeric(as.numeric(pred_model_logit1) ==
dat_use_hi$health_ins))

pred_model_logit1b <- (model_logit1$fitted > mean(dat_use_hi$health_ins))
table(pred_model_logit1b, dat_use_hi$health_ins)
frac_correct_l1b <- mean(as.numeric(as.numeric(pred_model_logit1b) ==
dat_use_hi$health_ins))

# examine how different cut-off values change predictive accuracy
frac_correct_try <- rep(0,18)
for (indx in 1:18) {
  pred_model_indx <- (model_logit1$fitted > (indx/20)  )
  frac_correct_try[indx] <- mean(as.numeric(as.numeric(pred_model_indx) ==
dat_use_hi$health_ins))
}
plot((seq(18)/20),frac_correct_try)
```

Also note that the code as given treats either miss (whether actually true and predict false, or actually false and predict true) as equally bad. In many applications this is not the case! Depending on the purpose of the model, false negatives and false positives could have different costs.

- Details of estimation

- recall that OLS just gives a convenient formula for finding the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ that minimize the sum

$$\sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki} \right) \right)^2 .$$ If we didn't know the formulas we could just have a computer pick values until it found the ones that made that squared term the smallest.

- similarly a probit or logit coefficient estimates are finding the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ that minimize

$$\sum_{i=1}^{n} \left(Y_i - f \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki} \right) \right)^2 ,$$ whether the $f(\square)$ function is a normal c.d.f. or a logit c.d.f.

- Maximum Likelihood (ML) is a more sophisticated way to find these coefficient estimates – better than just guessing randomly.

- For example the likelihood of any particular value from a normal distribution is the p.d.f., $\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. If we have

2 independent observations, X_1, X_2 from a distribution that is known to be normally distributed with variance of 1 (to keep the math easy) then the joint likelihood is $\dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(X_1-\mu)^2}\cdot\dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(X_2-\mu)^2}$. We want to find a value of μ that maximizes that function. This is an ugly function but we could note that any value of μ that maximizes the natural log of that function will also maximize the function itself (since $\ln(\ )$ is monotonic) so we take logs to get

$\ln\left(\dfrac{1}{\sqrt{2\pi}}\cdot\dfrac{1}{\sqrt{2\pi}}\right)-\dfrac{1}{2}(X_1-\mu)^2-\dfrac{1}{2}(X_2-\mu)^2$. Take the derivative with respect to μ and set it equal to zero to get

$(X_1-\mu)+(X_2-\mu)=0$ so that $\mu=\dfrac{(X_1+X_2)}{2}$. You should be able to see that starting with n observations would

get us $\mu=\dfrac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}$ so the average is also the maximum-likelihood estimator. A maximum-likelihood estimator could be similarly derived in cases where we don't know the variance (interestingly, that ML estimator of the standard error divides by *n* not *(n – 1)* so it is biased but consistent).

- Maximizing the likelihood of the probit model is one or two steps more complicated but not different conceptually. Having a likelihood function with a first and second derivative makes finding a maximum much easier than the random hunt.

Properly Interpreting Coefficient Estimates:

Since the slope, $\dfrac{\Delta Y}{\Delta X}=\dfrac{\Delta\Pr}{\Delta X}$, the change in probability per change in X-variable, is always changing, the simple coefficients of the linear model cannot be interpreted as the slope, as we did in the OLS model. (Just like when we added a squared term, the interpretation of the slope got more complicated.)

Return to the picture to make this clearer:
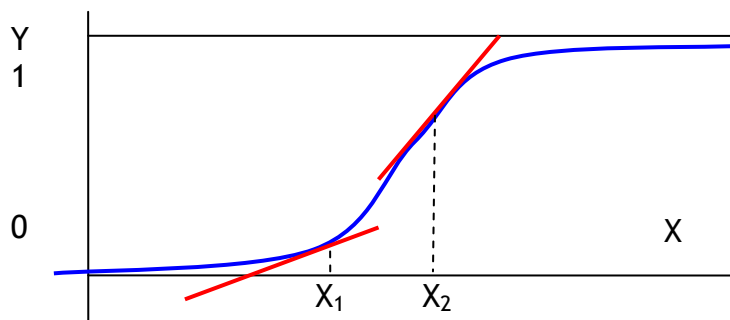
The slope at X_1 is rather low; the slope at X_2 is much steeper.

The effect of the coefficients now interacts with all of the other variables in the model: for example the effect of a person's gender on their probability of having health insurance will depend on other factors like their age and educational level. Women are generally less likely to have their own insurance than men, but

how much less? Among young people with very low education, neither men nor women are very likely to be insured; among older people with very high education both are very likely insured. The biggest difference is toward the middle.

For example, very simple logit and probit estimations on the CPS 2013 dataset (R program shows this in detail) gives the following coefficient estimates (I am suppressing notation on significance since it is not important here):

	coefficient estimates	
	logit	probit
(Intercept)	-0.37783	-0.2473
Age	0.002625	0.002951
I(Age^2)	0.000133	0.000057
female	-0.13458	-0.07423
AfAm	-0.49067	-0.2879
Asian	0.295029	0.1695
Amindian	-0.68546	-0.4059
race_oth	-0.1998	-0.1172
Hispanic	-0.40528	-0.2429
educ_hs	0.84353	0.5237
educ_smcoll	1.215126	0.7426
educ_as	1.54497	0.9321
educ_bach	2.146008	1.254
educ_adv	2.536002	1.444
married	0.602157	0.3499
divwidsep	-0.16488	-0.09745
union_m	1.407863	0.7217
veteran	-0.18023	-0.1157
immigrant	-0.68214	-0.3973
immig2gen	0.071965	0.03768

The probability of having health insurance varies for different socioeconomic groups. We can interpret the signs in a straightforward way: the negative coefficients on the "female" variable indicate that women are less likely to have health insurance. African-Americans are less likely, along with Hispanics and Native Americans. Educational qualifications are positive and get larger.

But how large are these differences? For example, how much less likely to have health insurance are immigrants? It depends on the other variables. Intuitively, if a person is male, highly-educated, and married then he's probably insured (being an immigrant would him only slightly less so). So the change in probability associated with immigrant status would be low. At the opposite end, a woman without a high school diploma, who is single, is already be unlikely to be insured. Immigrant status hardly changes this. Only in the middle will there be a big effect.

We can calculate it straightforwardly, though.

Consider, say, a 30-yr-old non-immigrant African-American woman with an advanced degree, whose predicted probability of having health insurance is

$$= f \begin{pmatrix} \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Female + \beta_4 Afam + \\ \beta_5 Asia + \beta_6 NativeAm + \beta_7 RaceOth + \beta_8 Hisp + \\ \beta_9 EdHS + \beta_{10} EdSmC + \beta_{11} EdAS + \beta_{12} Ed4 + \beta_{13} EdAdv \\ + \beta_{14} Marr + \beta_{15} DivWidSep + \beta_{16} Union + \beta_{17} Vet \\ + \beta_{18} Immig + \beta_{19} Imm2g + e \end{pmatrix}$$

$$= f \begin{pmatrix} \beta_0 \cdot 1 + \beta_1 \cdot 30 + \beta_2 \cdot 30^2 + \beta_3 \cdot 1 + \beta_4 \cdot 1 + 0 + 0 \\ + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 \\ + \beta_{16} \cdot 0 + \beta_{17} \cdot 0 + \beta_{18} \cdot 0 + \beta_{19} \cdot 0 + \beta_{20} \cdot 0 + \beta_{13} \cdot 1 \\ + 0 \ldots \end{pmatrix}$$

Summing the relevant coefficients (the intercept, female, and an advanced degree) gives a logit probability of

$$= f(-.378 + .079 + .120 - .135 - .491 + 2.536)$$

$$= \frac{1}{1 + e^{-(-.378 + .079 + .120 - .135 - .491 + 2.536)}}$$

Which is 85.0%. For an otherwise-identical immigrant woman (also with an advanced degree) the probability is 0.74, so the change in probability is about 11 percentage points.

Comparing the probit estimates, we would just change the functional form and use the normal cdf instead of the logit function, so again from:

$$= f \begin{pmatrix} \beta_0 \cdot 1 + \beta_1 \cdot 30 + \beta_2 \cdot 30^2 + \beta_3 \cdot 1 + \beta_4 \cdot 1 + 0 + 0 \\ + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 \\ + \beta_{16} \cdot 0 + \beta_{17} \cdot 0 + \beta_{18} \cdot 0 + \beta_{19} \cdot 0 + \beta_{20} \cdot 0 + \beta_{13} \cdot 1 \\ + 0 \ldots \end{pmatrix}$$

$$= f(-.247 + .089 + .051 - .074 - .288 + 1.444)$$

$$= pnorm(-.247 + .089 + .051 - .074 - .288 + 1.444) \text{ (in R)}$$

and find a probability for a non-immigrant woman as 0.835 and the immigrant woman to be 0.718, with a difference of 11.7 percentage points. These estimates from the logit and probit are very close.

Compare the change in probabilities for a divorced 45-yr-old white male without any degree, who is either an immigrant or not. Now the probability of having insurance is, by the logit, 0.461 for the non-immigrant and 0.302 for the immigrant, a change of 15.9 percentage points. From the probit the estimated probabilities are 0.462 for the non-immigrant and 0.311 for the immigrant, a change of 15.1 percentage points. This is because such a person is already less likely to have health insurance, so the difference of being an immigrant or not makes a bigger difference compared with the previous example.

Other Specifications

There are lots of other models that can be easily estimated – one of the advantages of R is that it makes it quite simple to use the same basic format of model specification but with different models. Some of these are just beginning to become more common in economics research.

You might in the future hear someone tell you, "you should try the *effin_magic* procedure" – and these notes will go through a series of "effin_magic" options but there are always more! But the general procedure for

whatever fancy effin_magic you're doing is to find a package in R that implements that effin_magic, split your data into training set and test set, do the effin_magic on the training set, then look at how it performs on the held-out test data. In time series, that test data might be going back a year to ask what if you had estimated a procedure from data up to last year then tried to predict one more year. In cross-section the selection of test data is often just random. But the basic idea is similar. There is often a tradeoff that an estimation can overfit the training set but then underfit the test set. (There are some procedures that use cross validation to do the same leave-some-out procedure on training data, to tune a parameter – it's turtles all the way down.) Much of the art comes from dealing with the curse of dimensionality. Now of course I encourage you to learn details of whatever effin_magic you're estimating, to read background literature on what other people have figured out, and as you gain experience you will get a better sense of which ones are likely to be best – but you can learn a lot by just trying them out.

Quantile Regression

If you recall our discussion of heteroskedasticity in things like the Age-Wage relationship, there is a well-known tendency for younger workers to have more compressed earnings, which then fan out as people get older.

For example, if we use the 2013 CPS data, we can look at people aged 25-55 who are working full time for most of the year and, even if we focus on a single educational group, for example those with a 4-year degree, we can see the spread here:
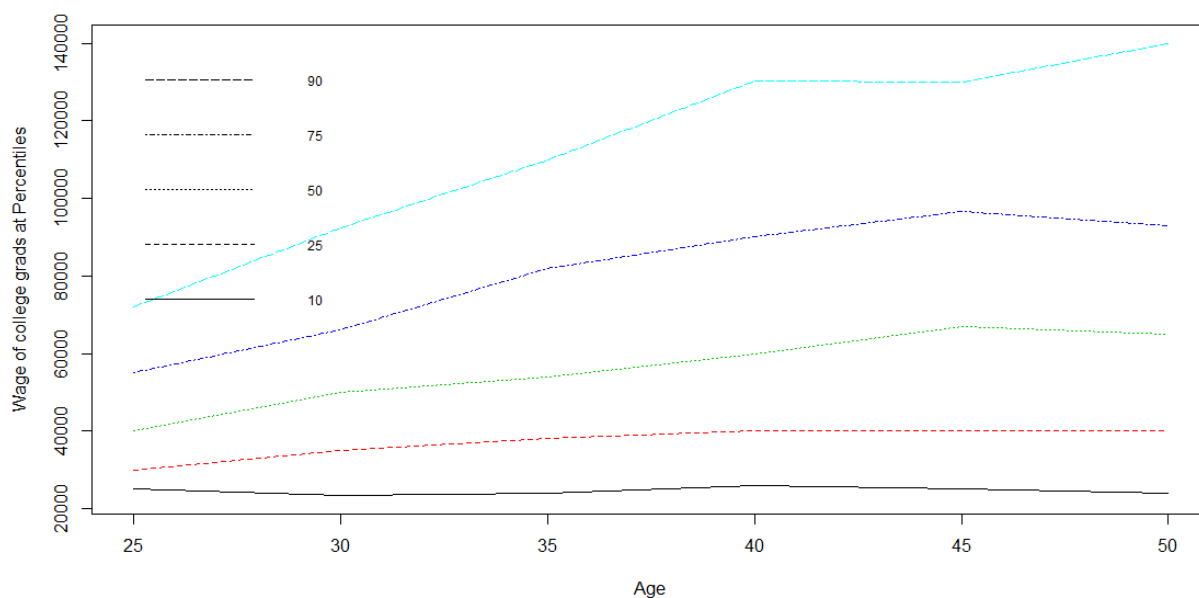
So the median worker saw a steady rise in wage: 30-yr-olds made \$50,000 while 50-yr-olds made about \$65,000; but those in the 25th percentile went from \$35,000 at age 30 to \$40,000 by 50; those in the 75th percentile went from \$66,000 to \$93,000.

One way to model these different results, for different percentiles, is with a quantile regression (mostly due to Roger Koenker), which uses a familiar regression framework to explain various percentiles.

In R this couldn't be easier: just use the "quantreg" package and call the `rq()` function instead of `lm()`. (Note that it's `rq` not `qr`; if you've done linear algebra you'll recall the QR matrix decomposition.)

```
p_tiles <- c(0.1, 0.25, 0.5, 0.75, 0.9)
```

```
quantreg1 <- rq(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian +
    race_oth + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach +
    educ_adv + married + divwidsep + union_m + veteran + immigrant +
    immig2gen,  tau=p_tiles, data = dat_use)
summary(quantreg1)
plot(quantreg1)
```

Details are in the R file, cps3.R. This estimates age-wage profiles like this (for women with a 4-year degree):
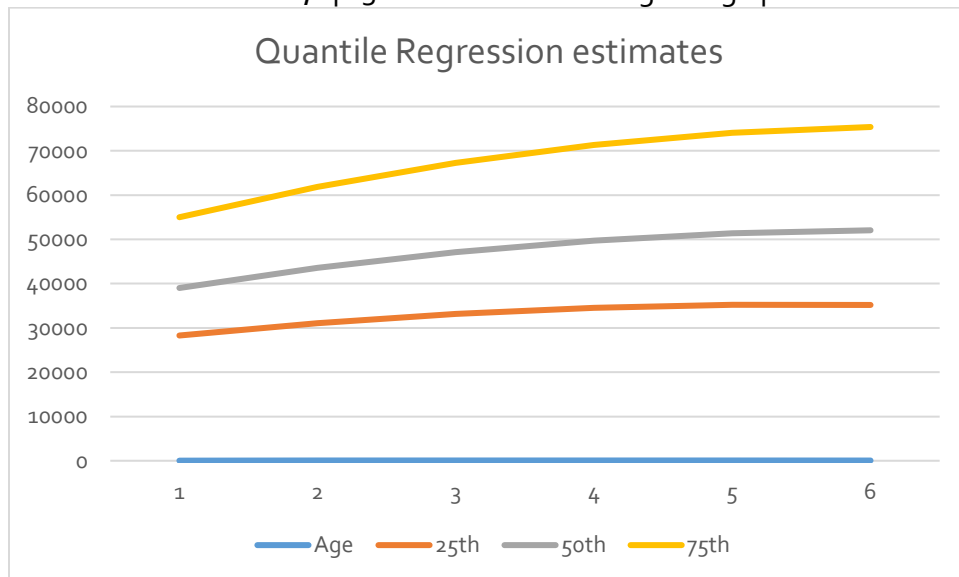
Which shows the spread.

Non-Parametric Regression

Instead of assuming a functional form – that the age-wage profile is linear, or quadratic, or cubic, or whatever … just let the data determine the wiggles in the function.

Peek at the underlying data, this is pdf of wages earned by different ages (those with college degree aged 25, 30, 35…):

Details in R program.

```
restrict2 <- as.logical(dat_use$educ_bach)

data3 <- subset(dat_use, restrict2)

NN <- length(data3$WSAL_VAL)

restrict3 <- as.logical(round(runif(NN,min=0,max=0.75)))

data4 <- subset(data3, restrict3)

library(np)

# note that this is rather computationally intensive!

model_nonparametric1 <- npreg(WSAL_VAL ~ Age, regtype = "ll", bwmethod =
        "cv.aic", gradients = TRUE, data = data4)

summary(model_nonparametric1)

npsigtest(model_nonparametric1)

plot(data4$Age,data4$WSAL_VAL, xlab = "age", ylab = "wage", cex=.1)

lines(data4$Age, fitted(model_parametric1), lty = 2, col = "red")

lines(data4$Age, fitted(model_nonparametric1), lty = 1, col = "blue")
```

A linear regression gives the expected value of Y given the values of X, under restriction that this expected value is a linear function. Quantile regression gives expected quantile of Y given X (again as a linear function). Nonparametric regression gives expected value of Y given X, subject to smoothness constraint (not linearity but still something).

LOESS

LOESS (local estimation with polynomials, not the kind of soil!) is related to nonparametric regression – where we think there is some smooth function y = f(x) but we want to estimate a very generic function, f( ). Unlike the nonparametric estimation previously it is much less computationally intensive (so runs much faster). The main limitation for our purposes is that X can have at most 4 variables, which must all be continuous (applies to R not in general).

```
model_loess1 <- loess(WSAL_VAL ~ Age,data3)
y_loess1_pred <- predict(model_loess1, data.frame(Age = seq(25, 55, 1)), se
        = TRUE)
plot(seq(25, 55, 1),y_loess1_pred$fit)
```

If you're starting to enjoy this stuff, I can recommend the (oddly titled) text, *Advanced Data Analysis from an Elementary Point of View*, by Cosma Shalizi. It is a terrific overview of these (and many more!) statistical techniques, with lots of examples in R, and great intuition for how the models work.

Spline & Generalized Additive Models

Again similar way of thinking of giving flexibility to functional form with tradeoff that more data and more computing power is needed. Estimate higher-order polynomials on sub-sections of the data, which come together at "knots". With one knot, this means cutting data into 2 subgroups; 3 knots gives 4 subgroups, etc. There are tuning parameters that control how many knots. (In R, the function smooth.spline with cv=TRUE.) Generalized Additive Models go farther along that route, allowing polynomials in various X variables; in R use the "gam" library.

Propensity Score Models

Again Gelman & Hill give a nice explanation. Ordinarily we look at estimating dummy variable coefficients using the whole set of data, so we want to estimate the coefficient on D in the equation, $y = X\beta + \gamma D + \varepsilon$ (where $X\beta$ includes all of the rest of the model variables). If the X variables are very similar for those with D=0 and D=1, then we are likely to get a good estimate of the effect of D (the γ coefficient). But if the values of the of X variables are very different, between those with D=0 and those with D=1, then we need to be sure that the model is very accurate.

As a simple example, consider again the sort of model we'd discussed about dummy variables – suppose we want to estimate something like this model,
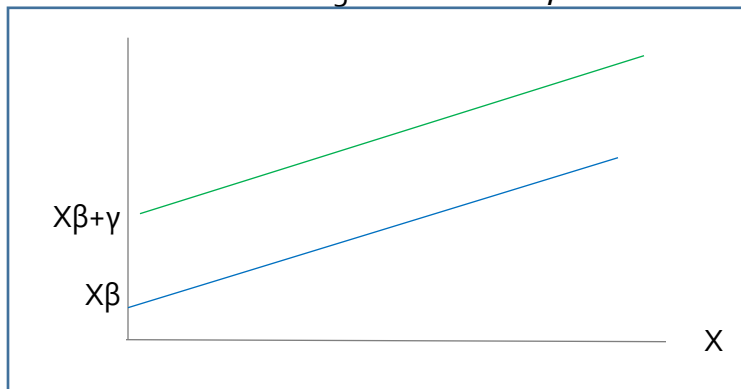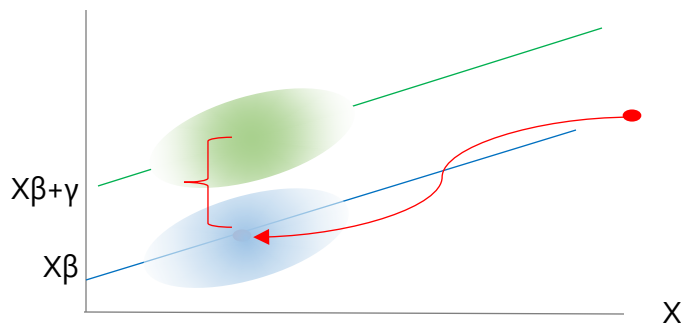

If the data for D=1 and D=0 are similar, then this can be well estimated:
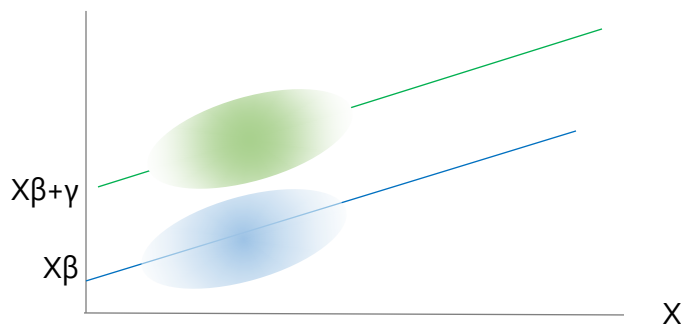
If, however, we consider how to use a point such as this one:
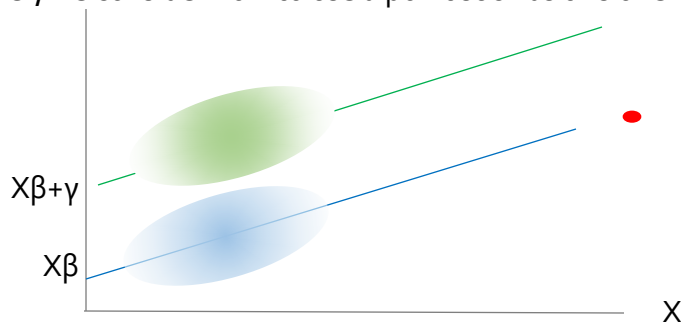
Then what the model is essentially doing is using the estimate of β to shift that down to a comparable location then measuring the vertical distance, so:

But what if the estimate of β is a bit off? What if, instead of a simple linear function like Xβ, we have some nonlinear part? Or an interaction of X and D that is omitted? In that case the new point might be just contributing noise.

So a propensity score model would just compare D=1 values with those certain D=0 values that have X-values that are "close" – leaving out the X-values that are far away. If X is uni-dimensional then defining "close" is pretty easy (as in the graph above) but if X has multiple dimensions then this becomes more difficult – recall our discussion of k-nearest-neighbor for machine learning!

To do this in R, start with a logit model of the 'treatment' – which for this example is whether the person is female. Then use this estimated distance to match.

```
model1 <- glm(female ~ Age + I(Age^2) + AfAm + Asian + Amindian + race_oth

    + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach + educ_adv

    + married + divwidsep + union_m + veteran + immigrant + immig2gen,

    family= binomial, data = dat_use)

X_dist <- model1$fitted

Y_est <- dat_use$WSAL_VAL

tr_est <- dat_use$female


require('Matching')

# this is numerically intensive

model_match <- Match(Y=Y_est, Tr=tr_est, X=X_dist, M=1, version='fast')

summary(model_match)
```

This estimates the female wage disadvantage to be -18776, compared to a linear regression model where the dummy variable gets an estimate of -19296, so not much of a difference in this case, although other situations might find a bigger difference in estimates.

Alternately we could consider education, which is a bit more of a "treatment" and look at the effect of getting an advanced degree compared with getting a bachelor's degree.

```
use_varb2 <- as.logical(dat_use$educ_bach + dat_use$educ_adv)

dat_use2 <- subset(dat_use,use_varb2) # 19231 obs
```

```
model2 <- glm(educ_adv ~ Age + I(Age^2) + female + AfAm + Asian + Amindian +
        race_oth + Hispanic + married + divwidsep + union_m + veteran +
        immigrant + immig2gen, family= binomial, data = dat_use2)

X_dist <- model2$fitted

Y_est <- dat_use2$WSAL_VAL

tr_est <- dat_use2$educ_adv



require('Matching')

model_match2 <- Match(Y=Y_est, Tr=tr_est, X=X_dist, M=1, version='fast')

summary(model_match2)



modelcompare2 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + AfAm + Asian +
        Amindian + race_oth + Hispanic + educ_adv + married + divwidsep +
        union_m + veteran + immigrant + immig2gen, data = dat_use2)

summary(modelcompare2)
```

Here's a good explanation, http://ftp.iza.org/dp1588.pdf

Lasso

Lasso (and Spike and Slab, below) are both used for selecting which variables are "important" in predicting. Note as usual that important in prediction might not be the same as causal, however again we can explore the data to see. Both techniques will pare off X-variables that do not contribute much predictive value to the regression. In cases where we have very few observations (i.e. most of macro), these would not be appropriate, however in cases with dense data then it is reasonable to consider – if your variable of interest is not selected for prediction, then you have to think about why.

Much of the impetus for developing these sorts of models comes from either websites (that get arrays of data streaming through, and try to figure out which have any predictive value) or genomics (which have huge numbers of candidate genetic markers, and try to figure out which have predictive value).

Lasso is Least Absolute Shrinkage and Selection Operator, and in R is usually implemented with the `lars` package.

This finds coefficients that not only minimize the squared residuals (just like OLS) but also tries to minimize the squared coefficient sizes – so it penalizes 'too many' explanatory variables. In machine learning this is a way of finding efficient predictors but for our purposes it helps to see which variables are important in the model.

It is useful to ensure that the X-variables are scaled similarly; this will do the trick.

```
x_varb <- cbind(Age,I(Age^2), female, AfAm, Asian, Amindian,race_oth,
        Hispanic, educ_hs, educ_smcoll, educ_as, educ_bach, educ_adv,
        married, divwidsep, union_m, veteran, immigrant, immig2gen)
stand_Z <- function(X) {
  rval <- matrix(data = NA, nrow = nrow(X), ncol = ncol(X))
  for(j in 1:ncol(X)) rval[,j] <- (X[,j] - mean(X[,j]))/sd(X[,j])
  return(rval)
  }
x_varb_dm <- stand_Z(x_varb)
dimnames(x_varb_dm) <- dimnames(x_varb)
require(lars)
```

```
model_lars <- lars(x_varb_dm,WSAL_VAL)
summary(model_lars)
plot(model_lars)
coef(model_lars)
```

We can get an idea of how it classifies the importance of the different factors from our basic wage regression,

<insert examples of output here>

Related to Lasso Regression is the Ridge Regression (for cases with near multicollinearity) and Elastic Net Regression (which combines them). Lasso might cut off too much so Elastic Net can give a bit more ... elasticity. It is often important to regularize the x-variables so either get them to be mean-zero and stdev=1 or to be on [0,1] interval. Try the `glmnet` package, so if its alpha parameter is set to zero you have ridge; if alpha is 1 then lasso; in between is elastic net.

Spike & Slab

There are many other regression techniques.

Spike and Slab (the name refers to the Bayesian prior distributions about coefficients) is implemented in R with the `spikeslab` package. Scott and Varian (2012) refer to the "fat regression" problem where there are more possible explanatory variables than there are observations – there is a severe problem with degrees of freedom. The "spike" refers to the probability that a particular variable is in the model (there is either a 0 or a 1 to select that particular explanatory variable) while the "slab" is the information from the coefficient estimates. (The LASSO estimator also approached this problem.)

This is another way to gauge the importance of various parts of your model, particularly in cases if there are lots of interactions.

A linear regression with a lot of interactions (returning to our usual CPS wage regression) could include this,

```
modelcompare <- lm(WSAL_VAL ~ (Age + I(Age^2) + female + AfAm + Asian +
    Amindian + race_oth + Hispanic + educ_hs + educ_smcoll + educ_as +
    educ_bach + educ_adv + married + divwidsep + union_m + veteran +
    immigrant + immig2gen) ^2 + (industry_f + occupatn_f +
    state_f)*female, data = dat_8)
summary(modelcompare)
```

Whereas a version with spike and slab would use this code,

```
require(spikeslab)
set.seed(54321)
model1_spikeslab <- spikeslab(WSAL_VAL ~ (Age + I(Age^2) + female + AfAm +
    Asian + Amindian + race_oth + Hispanic + educ_hs + educ_smcoll +
    educ_as + educ_bach + educ_adv + married + divwidsep + union_m +
    veteran + immigrant + immig2gen) ^2 + (industry_f + occupatn_f +
    state_f)*female,  data = dat_8)
summary(model1_spikeslab)
print(model1_spikeslab)
plot(model1_spikeslab)
```

Both will keep your computer running for a while! Note the "set.seed" sets the random number generator so that, if you try it again, you'll get the same output as I did.

The picture is tough to interpret given so many lines,
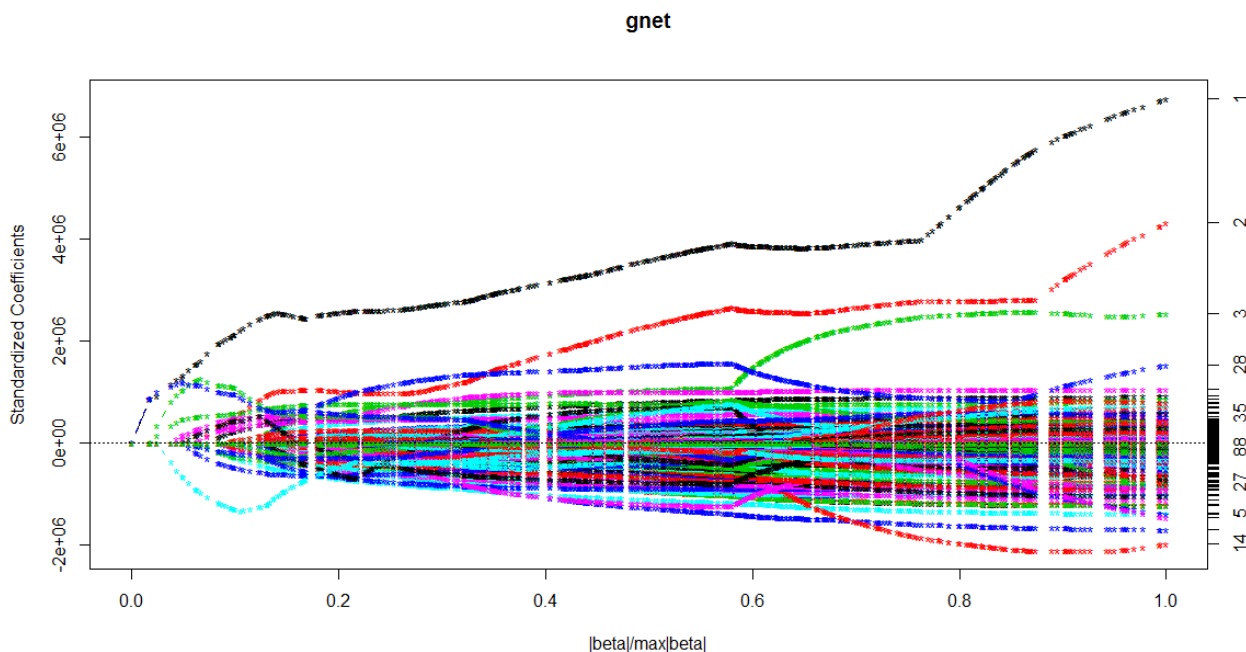
gnet

Other than that there are only a few that really stand out. The "print" call will give the coefficient estimates from this model; the top of that print is:

---> Top variables:

	bma	gnet	bma.scale	gnet.scale
Age:educ_adv	16690.17	17715.13	1123.054	1192.021
Age:educ_bach	8792.73	11872.62	485.761	655.912
Age	7020.34	7152.103	817.248	832.587
occupatn_f17	-6143.29	-6595.35	-18991.2	-20388.7
occupatn_f8	-5395.99	-5487.32	-23699.6	-24100.8
occupatn_f10	4805.165	4645.705	20435.86	19757.69
occupatn_f6	-4621.3	-4926.29	-33460.7	-35668.9
occupatn_f21	-4546.55	-4899.63	-18791.1	-20250.4
occupatn_f22	-4533.39	-4921.73	-19702.5	-21390.3
female:occupatn_f10	-4424.35	-4457.37	-22250	-22416.1

Where the "bma" (Bayesian Model Averaging) and "gnet" (the generalized elastic net, with penalty parameters for coefficients) refer to different estimation methods; the first two columns are coefficients for the normalized values of the x-variables (with mean 0 and std dev 1) while the last two columns are the usual coefficient estimates.

From looking at the top ones most likely to be selected for inclusion in the model, we see that the first 2 most important variables are age interacted with education measures, then age, then various occupation categories. This is similar to the LASSO that implied that education was most important.

(If you learn nothing else from this course, learn that the data show that education is important! Although, you know, probably because people with more education actually learn and remember the s*** that their professors say…)

Estimation with Trees and Forests

With Tree Models (from computer science) the emphasis is on prediction not necessarily causation. This can make economists crazy although it can also be a good way to get at causation – are there certain "features" (which is the term that computer science uses instead of 'explanatory variables') that can easily classify some outcome? This can be part of a data description or modeling exercise.

A tree model gives a series of splits of the X-variables (which, just like the Y-variable, might be, but need not be, discrete) in order to subdivide and subdivide. It's a good way of handling models where there are crazy degrees of interaction, where some regions of the X-variables imply very different Y-behaviors than other X-regions.

The tree is constructed by finding the split in an X-variable that most reduces the sum of squared errors in each stem of the tree (if Y is continuous). Since we have been talking about "sum of squared errors" since OLS models, this should be reassuring to you.

An R program to predict whether a person is covered by employer-provided health insurance is:

```
library('rpart')
# tree model of whether has health insurance
model1 <- rpart(health_ins ~ Age + I(Age^2) + female + AfAm + Asian +
      Amindian + race_oth + Hispanic + educ_hs + educ_smcoll + educ_as +
      educ_bach + educ_adv + married + divwidsep + union_m + veteran +
      immigrant + immig2gen,data = dat_use_hi, method = "class")
summary(model1)
# plot(model1)
# text(model1, use.n = TRUE, all=TRUE, cex=.8)
post(model1, file = "tree_1.ps",
      title = "Classification Tree for Health Insurance")
```

Note that since the y-variable is 0/1, it uses "`method = "class"`, whereas if the y-variable were continuous it would use "`method = "anova"`.

Classification Tree for Health Insurance
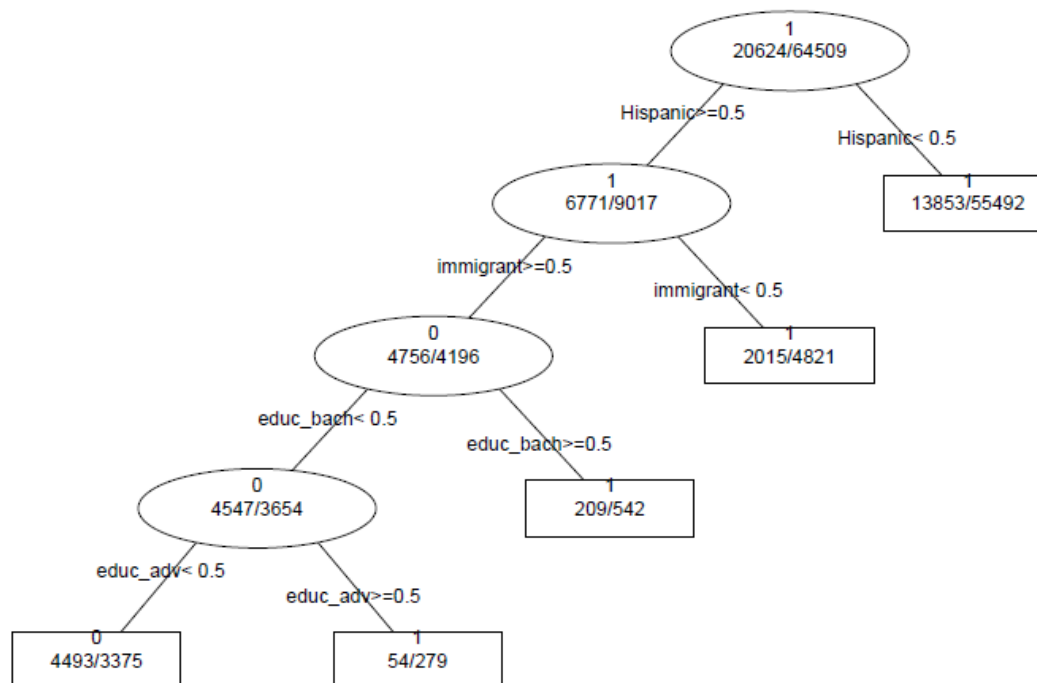
We could improve this method by going back to the idea that we discussed with k-nn, where we split into training and evaluation sets – use 80% of the data to train the tree, then see how well it would classify the remaining 20%. This helps if you worry about overfitting. There are other methods of pruning trees to keep them from growing too much.

But note that this demolishes the idea of "statistical significance". Various econometricians have developed methods that would take a set of X variables and search over them to figure out the "best" ones to explain the variation in Y (where "best" is usually something like R² but penalized for complexity and number of variables in the model). That's the same basic idea except that we cannot then go on to cite p-values. If I

create a regression and examine a p-value then it has some information about how likely it is, that I'm being fooled and could see such a coefficient just by chance. But if I'm p-hacking (finding the regression with the lowest possible p-value) then I need to be asking more sophisticated versions of "am I being fooled by randomness".

Trees and Forests

Next we can go from creating a single tree to growing a whole forest.

Random Forests are more complex although they can offer improvements to classification accuracy. They are notoriously difficult to understand or explain, however – they are often mostly a "black box". (Cathy O'Neil has book on policy implications of such.) Nevertheless they can be a useful method of classification even if as a comparison – if a random forest model classifies A% correctly while your preferred model gets B%, then the difference (A-B) can be a useful way to assess how good is the model.

The idea of a Random Forest is to take a randomly-chosen sub-set of the data and build a tree model from it. Then take another randomly-chosen sub-set and build another tree. And another and another... Take these trees and aggregate them (perhaps build 10 trees and figure out if 7 imply one outcome whereas 3 imply the other outcome).

```
# random Forest
library('randomForest')
set.seed(54321)

# the command system.time() tells how long it takes
system.time(model3 <- randomForest(as.factor(health_ins) ~ .,
      data=dat_cps_rf, importance=TRUE, proximity=TRUE))
print(model3)
round(importance(model3),2)
varImpPlot(model3)
```

The Random Forest gives a "Confusion matrix" comparing the ones that are truly 0/1 versus what is predicted:

	actual 0	actual 1
predicted 0	1558	1961
predicted 1	806	9988

The previous logit model gives results of:

	actual 0	actual 1
predicted 0	5363	3415
predicted 1	15261	61094

The numbers of observations are different because I had clipped the size of the data for the random forest in order to economize on computing time. So it's not apples-to-apples but skewed in favor of logit (since that's got more information). But if we look at the fraction in each class, we see that:

	random forest		logit model	
	actual 0	actual 1	actual 0	actual 1
predicted 0	0.109	0.137	0.063	0.040
predicted 1	0.056	0.698	0.179	0.718

So the random forest mis-classified 19.3% of the observations while the logit model mis-classified 21.9% - so even with nearly six times more observations, the logit was a worse fit overall. (You can tweak both methods to do better, maybe a forest of conditional inference trees would be better or you can better specify the logit. These results are illustrative.)

Random forests can also be done for regression problems – the dependent variable need not be 0/1 as above but can be a continuous variable.

In classification problems, there are a range of options: logit or probit (binary or multinomial), trees/forests, then Support Vector Machines. There's no obvious best option so play around to see!

These methods are still relatively new in economics; see Hal Varian's piece on *Big Data: New Tricks for Econometrics*.

Support Vector Machines

Another method of classifying data, usually when Y is 0/1 (or a limited number of outcomes). Package is `e1071`. These depend on tuning parameters with usual tradeoff between variance and bias – better performance in training samples can mean worse performance in test samples. It is related to logistic regression and can give similar performance. Chapter 9 of James, Witten, Hastie & Tibshirani is a nice explanation.

Factor Analysis

Another common procedure, particularly in finance, is a factor analysis. This asks whether a variety of different variables can be well explained by common factors. Sometimes when it's not clear about the direction of causality, or where the modeler does not want to impose an assumption of causality, this can be a way to express how much variation is common. As an example, one price that people often see, which changes very often, is the price of gasoline. If you have data on the prices at different gas stations over a long period of time, you would basically see that while the prices are not identical, they move together over time. This is not surprising since the price of oil fluctuates. There might be interesting variation that at some times certain stations might be more or less responsive to price changes – but overall the story would be that there is a common influence.

Factor Analysis (and the related technique of Principal Components Analysis, PCA) are not model-based and can be useful methods of exploration. An example might be the easiest way to see how it works.

I got daily data from Federal Reserve on Eurodollar interest rates for 1-, 3-, and 6-months, from 1971-2014 (so called since it was originally the rate to borrow dollars from a bank in London, which remains the center of this market).

```
prcomp1 <- prcomp(~ ed1m + ed3m + ed6m, data = data_2)
summary(prcomp1)
```

Which shows that the first principal component explains 99.7% of the variation in these interest rates.

(With a wider span of maturities, we often find that 3 factors explain most interest rate movements: level, slope, and curvature.)

Prediction and Causality

So to return to the generic example, "effin_magic", if your goal is simply prediction then you might estimate a series of different models -- effin_magic1, effin_magic2, ... For each one, you would split the data into training and test sets, estimate the effin_magic on the training set, then evaluate how well it does on the test set. For some estimation procedures the training set will be again split so that the model can tune hyperparameters. You repeat the estimation a number of times for different randomly-chosen test sets, to evaluate how robust the results are. Then do this for the next effin_magic and the next and the next. When evaluating predictions, your measure of goodness of fit can change depending on the problem. You should also look at goodness of fit for different subsets of data – does it do much better or worse for certain groups? (For example, a marketing algo might change around the holiday season when people are perhaps less likely to be shopping for themselves and more likely to be looking for gifts.) You might end up with an ensemble of models so for certain groups or circumstances one model is best for prediction.

While prediction is often one criteria, often in economics we also want to consider causation. Getting from a good predictive model to a causal model is difficult and requires more theory. You will develop that as you go along, but pay attention to questions of endogeneity and omitted variables.

Experiments and Quasi-Experiments
- ideal: double-blind random sort into treatment and base sets
- differences estimator for "natural experiments" or quasi-experiments
- Problems can be internal:
    - incomplete randomization
    - failure to follow treatment protocol
    - attrition
    - experiment (Hawthorne) effects
- or external
    - non-representative sample
    - non-rep program
    - treatment/eligibility
    - general equilibrium effects

Time Series
Basic definitions:
- first difference $\varsigma Y_t = Y_t - Y_{t-1}$
- percent change is $\%\Delta Y_t = \dfrac{\Delta Y_t}{Y_{t-1}}$ and is approximately equal to $\ln(Y_t) - \ln(Y_{t-1})$ – this log approximation is commonly used
- lags: the first lag of Y_t is Y_{t-1}; second lag is Y_{t-2}, etc.; sometimes use lags of differences
- Autocorrelation: how strong is last period data related to this period? The autocorrelation coefficient is $\rho_j = \dfrac{\text{cov}\left(Y_t, Y_{t-j}\right)}{\text{var}\left(Y_t\right)}$ for each lag length, j. Sometimes plot a graph of the autocorrelation coefficients for various j.
- Common assumption: Stationarity: a model that explains Y doesn't change over time – the future is like the past, so there's some point to examining the past – a crucial assumption in forecasting! But this is why we usually use stock returns not stock price – the price is not likely stationary even if returns are. (Also often assume ergodic.)
- If autocorrelations are not zero, then OLS is not appropriate estimator if X and Y are both time series! The standard errors are a function of the autocorrelation terms so cannot properly evaluate the regression.
- Seasonality is basically a regression with seasons (months, days, whatever) as dummy variables. So could have $Y_t = \beta_0 + \beta_1 January + \beta_2 February + \beta_3 March + \ldots + \beta_{11} November + u_t$ - remember to leave one dummy variable out! Or $Y_t = \beta_0 + \beta_1 Monday + \beta_2 Tuesday + \ldots + \beta_{11} Saturday + u_t$.

Types of Models
- AR(1) – autoregression with lag 1
- $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$
- Forecast error is one-step-ahead error

- Note that can re-write the AR(1) equation, by substituting $Y_{t-1} = \beta_0 + \beta_1 Y_{t-2} + u_{t-1}$, as
$Y_t = \beta_0 + \beta_1(\beta_0 + \beta_1 Y_{t-2} + u_{t-1}) + u_t = \beta_0(1 + \beta_1) + \beta_1^2 Y_{t-2} + \beta_1 u_{t-1} + u_t$, then substitute in for
$Y_{t-2} = \beta_0 + \beta_1 Y_{t-3} + u_{t-2}$, and so on. So the current value is a function of all past error terms,
$Y_t = \beta_0(1 + \beta_1 + \beta_1^2 + \ldots + \beta_1^T) + [u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \ldots + \beta_1^T u_{t-T}] + \beta_1^T Y_{t-1-T}$. Note that as long as
$|\beta_1| < 1$, the last term drops and the sums converge as $T \to \infty$.

- Reminder of convergent series: look at $(1 + \beta_1 + \beta_1^2 + \ldots + \beta_1^T)$, note that
$\beta_1(1 + \beta_1 + \beta_1^2 + \ldots + \beta_1^T) = (\beta_1 + \beta_1^2 + \ldots + \beta_1^{T+1})$. Add and subtract β_1^{T+1} and fiddle the parentheses to
write $(1 + \beta_1 + \beta_1^2 + \ldots + \beta_1^T) = 1 + (\beta_1 + \beta_1^2 + \ldots + \beta_1^T + \beta_1^{T+1}) - \beta_1^{T+1}$. Notate that ugly term
$(1 + \beta_1 + \beta_1^2 + \ldots + \beta_1^T) = Z$, then the equation says that $Z = 1 + \beta_1 Z - \beta_1^{T+1}$. Solve,
$Z - \beta_1 Z = Z(1 - \beta_1) = 1 - \beta_1^{T+1}$, and $Z = \dfrac{1 - \beta_1^{T+1}}{1 - \beta_1}$. Substitute this into the previous equation for Y_t

- $Y_t = \beta_0 \dfrac{1 - \beta_1^{T+1}}{1 - \beta_1} + [u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \ldots + \beta_1^T u_{t-T}] + \beta_1^T Y_{t-T}$. As $T \to \infty$, the first term goes to $\beta_0 \dfrac{1}{1 - \beta_1}$

, the last term goes to zero, and the middle term is $\displaystyle\sum_{\tau=0}^{\infty} \beta_1^\tau u_{t-\tau}$.

- If $\beta_1 = 1$ then none of the terms converge – the model becomes a random walk or integrated with
order 1, I(1) or has a unit root. (Can test for this, most common is Augmented Dickey-Fuller ADF.)
  - Also random walk with trend, so $Y_t = \beta_0 + \gamma t + Y_{t-1} + \varepsilon$
  - And random walk with drift, so $Y_t = \beta_0 + Y_{t-1} + \varepsilon$ (but no trend)
  - Or just plain random walk, $Y_t = Y_{t-1} + \varepsilon$
- Random walk means that AR coefficients are biased toward zero, the t-statistics (and therefore p-values) are unreliable, and we can have a "spurious regression" – two time series that seem related only because both increase over time. Consider this case of variables X and Y, each of which are $Z_t = 1 + Z_{t-1} + \varepsilon$ where ε is a random draw from a normal distribution.

```
rm(list = ls(all = TRUE))

const_term <- 1
ar_coeff <- 1
start_val <- 100
num_terms <- 100

x_val <- matrix(data = NA, nrow = num_terms, ncol = 1)
y_val <- matrix(data = NA, nrow = num_terms, ncol = 1)

x_val[1] <- start_val
y_val[1] <- start_val

set.seed(12345)
x_rand <- rnorm(num_terms, mean = 0, sd = 1)
y_rand <- rnorm(num_terms, mean = 0, sd = 1)

for (indx in 2:num_terms) {
  x_val[indx] <- ar_coeff*x_val[indx - 1] + const_term + x_rand[indx]
  y_val[indx] <- ar_coeff*y_val[indx - 1] + const_term + y_rand[indx]
}

model1 <- lm(y_val ~ x_val)
```

```
summary(model1)

(ar(y_val)) #AR method
```

- AR(p) – autoregression with lag p
- $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + u_t$
-
- ADL(p,q) – autoregressive distributed lag model with p lags of dependent variable and q lags of an additional predictor, X.
- Need usual assumptions for this model
- Lag length? Some art; some science! Various criteria (AIC, BIC, given in text) to select lag length.
- Granger Causality – jargon meaning that X helps predict Y; more precisely X does not Granger-cause Y if X does not help predict Y. If X does not help predict Y then it cannot cause Y.
- Trends provide non-stationary models
- Random walk non-stationary model:
- Breaks can also give non-stationary models
- test for breaks, sup-Wald test
- Cointegration "The Definitive Overview", ftp://ftp.econ.au.dk/creates/rp/14/rp14_38.pdf
- Can model time series as regression of Y on X, of ln(Y) on ln(X), of ΔY on ΔX, or of %ΔY on %ΔX (where, recall, %$\Delta Y = \Delta \ln Y$ since the derivative of the log is the reciprocal) – this is where the art comes in!
- Distributed lag models can be complicated (Chapter 15) and so we want at a minimum Heteroskedasticy and Autocorrelation Consistent (HAC) errors – like the heteroskedasticity-consistent errors before (Newey-West)
- VAR – Vector AutoRegression, incorporate k regressors and p lags so estimate as many as k*p coefficients – these are classic in macro modeling, following work of Chris Sims
- GARCH models – Generalized AutoRegressive Conditional Heteroskedasticity models – allow the variance of the error to change over time, depending on past errors – allows "storms" of volatility followed by quiet (low-variance)
  - $y_t = \sigma_t \varepsilon_t$; $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$ GARCH(p,q)
  - Combine with random walk analysis for IGARCH, etc

In R: read "Time Series Analysis with R" for a high-level overview of what's possible – that has refs to various packages that you can study, as you figure out what exactly you want to do.
http://www.stats.uwo.ca/faculty/aim/tsar/

If you fall in love with time series analysis, James Hamilton has a big textbook that can help

Methodology
As you get more experience with econometrics you can start to understand the old jokes about why the discipline name includes "con" and "trics"! Ed Leamer has a classic paper, *Let's Take the 'Con' Out of Econometrics*. Diedre McCloskey has been a persistent critic, e.g. in *Knowledge and Persuasion in Economics* or *The Trouble with Mathematics and Statistics in Economics*. Chris Sims wrote, *Why are Econometricians so Little Help?* Although Angrist and Pischke wrote *Mostly Harmless Econometrics*. You can understand why so many econometricians advise, "beware of econometricians."

More…

Econometrics goes on and on – there are thousands of techniques for new situations and new conditions, especially now that computing power quickly increases the amount of calculations that can be done. There is so much to learn!