Lecture Notes

PSY V0500, Statistical Methods in Psychology

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

Spring 2024

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 DataViz textbook and Hawkes online material and better correlated with videos – review them all, they're not substitutes.

Table of Contents

PSY V0500, Statistical Methods in Psychology 1
Intro and how to R 3
Know Your Data 4
The Challenge 4
Step One: Know Your Data 6
Show the Data 6
Histograms 6
Hawkes 10
Basic Concepts: Find the Center of the Data 11
Spread around the center 13
Now Do It! 16
On Correlations: Finding Relationships between Two Variables 17
Use Your Eyes 17
How can we measure the relationship? 19
Sample covariances and sample correlations 21
Important Questions 22

Intro and how to R

Welcome! This is Psych V0500 Statistical Methods in Psychology, here at the department of psychology, in the Colin Powell School at the City College. This is going to be fun! I love Stats, which is why the department asked me to teach this class. I hope to share my passion for the topic.

Statistics 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. Psychology is currently said to be in crisis from well-known studies not replicating – sometimes through outright fraud. The bottom line, for many researchers, is that "low p-value" leads to "publication". There might be a few who simply accept that causality; many more would defend a more cynical view that other people believe that so they have to play along. We'll talk about some of the rationale for why that view became so common over the past century of scientific research. Sometimes people's view of research reminds me of Bitcoin mining: in both, someone starts with a bunch of numbers, processes them through some series of computations, then if a certain pattern is discovered, that pattern has value – whether the pattern is a few low p-values or zeros in a hash. Researchers should be more concerned with understanding their data.

To understand data we're going to explore techniques in Data Visualization – thus the text. We'll use the book, *Data Visualization*, by Kieran Healy. You can buy the book (depending on your preferences regarding dead trees) but it's also available online for free and there is a lot of material on his github page. Data Viz requires a basic understanding of stats, thus the other Hawkes online text.

For Hawkes, most of that should be review so I'll only go over that material from a relatively high level. This class is meant for students who have a decent background in basic statistics. In my experience however, many students don't know if they have a decent background in basic stats – that's the whole problem. So there's a diagnostic test a few weeks into the semester that you take online. The online material also lets you review since I know many students haven't done stats in a while and have to do some mental archaeology. The diagnostic test will only count for 10% of your grade but if you do poorly then we should talk. You must purchase access to the online Hawkes Learning System in order to take the test. This system gives straightforward explanations of each topic and then allows you to do lots and lots of practice problems, which is the best way to ensure that you learn those topics. For that, go to https://learn.hawkeslearning.com. Enter your school (listed as "CUNY - The City College of New York") and find "Beginning Statistics". I have set up the Hawkes system with recommended practice sections but you can explore more. You can do it at your own pace – some might want to get an early start. Each chapter has practice problems. There are 2 additional practice exams, which you can take as many times as you'd like. Your score is recorded but doesn't count in the grade, it's just FYI. Practice #1 is a bit easier, Practice #2 is a bit tougher than the actual diagnostic test. You have only one chance to take the diagnostic test before the deadline so find a time when you have 2 hours to go through it. There is a strong correlation between doing lots of practice and performing well.

In this class we're going to be using a computer program called R, and we're going to use R Studio, which sits on top of R. The DataViz text gives some help on that too. One of the very first things you've got to do is install these bad boys -- so how do you do that? You go to that link that I just put up there and you follow the instructions. Step one: download and install R; step two: download and install R Studio desktop. I can't give exact instructions since this depends on your operating system and computer setup. I happen to be using a windows machine, you may have an apple machine, you may have some variety of linux , so the instructions are going to be variable for each person. I would love to give you step-by-step details but I cannot. So pause

here and try it. You can contact me if you're having trouble; I'm happy to get on zoom and help you through step by step. But I think most of you know how to install a program – or in this case, two separate programs, R and R Studio. Once those are installed you'll only need to open up R Studio and that will automatically open up R.

See Lecture 1 A and 1B plus videos for more.

Know Your Data

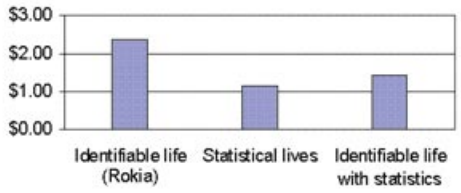
We begin with "Know Your Data" and "Show Your Data," to review some of the very initial components necessary for data analysis.

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.

Science is a natural fit for statistical analysis. 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?*"

Statistics 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.

Psychology is currently said to be in crisis from well-known studies not replicating – sometimes through outright fraud. The bottom line, for many researchers, is that "low p-value" leads to "publication". There might be a few who simply accept that causality; many more would defend a more cynical view that other people believe that so they have to play along. We'll talk about some of the rationale for why that view became so common over the past century of scientific research. Sometimes people's view of research reminds me of Bitcoin mining: in both, someone starts with a bunch of numbers, processes them through some series of computations, then if a certain pattern is discovered, that pattern has value – whether the pattern is a few low p-values or zeros in a hash. Researchers should be more concerned with understanding their data.

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! A Likert scale asking people to rate feelings on scale of 1-5 is different from a scale rating from 1-10.

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.

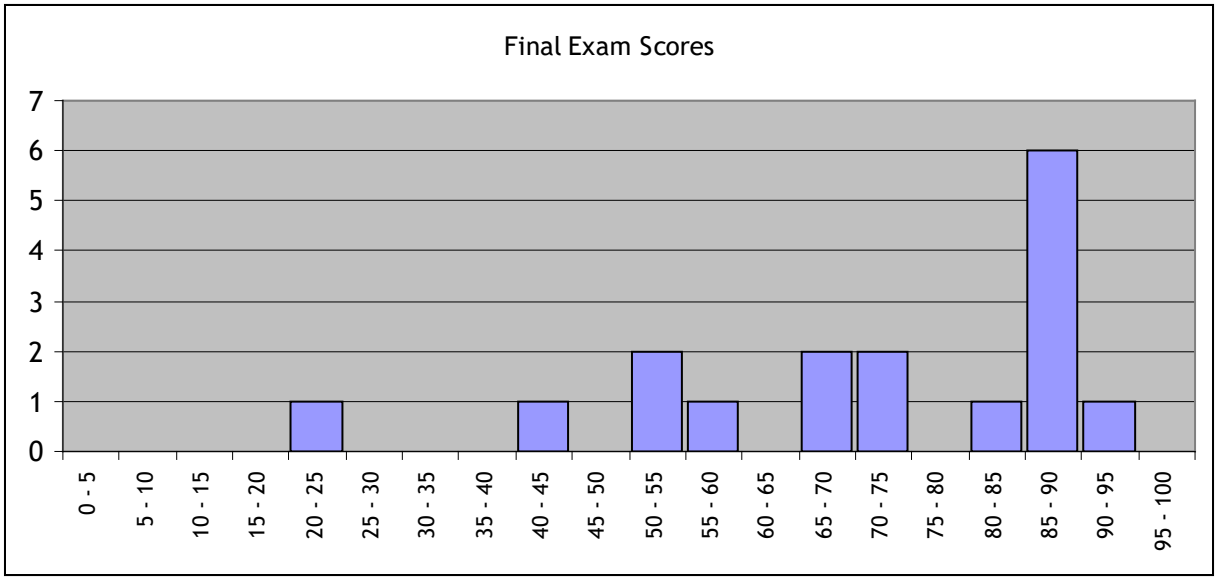
We're drowning in data because increasing computing power makes so much more available to us. The data piles up while nobody's looking at it. An online store might generate data on the thousands of clicks simultaneously occurring, but it's probably just spooling onto some server's disk drive. It's just 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 African plains, watching all around ourselves to find food 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. The homework will ask you to try your hand at it.

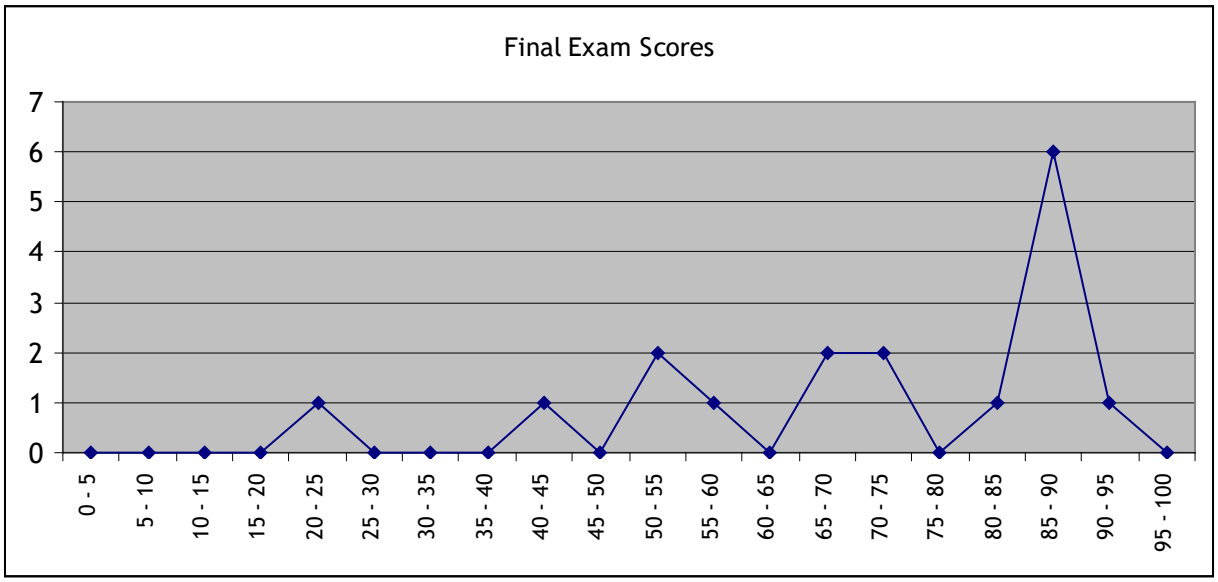
Histograms

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:

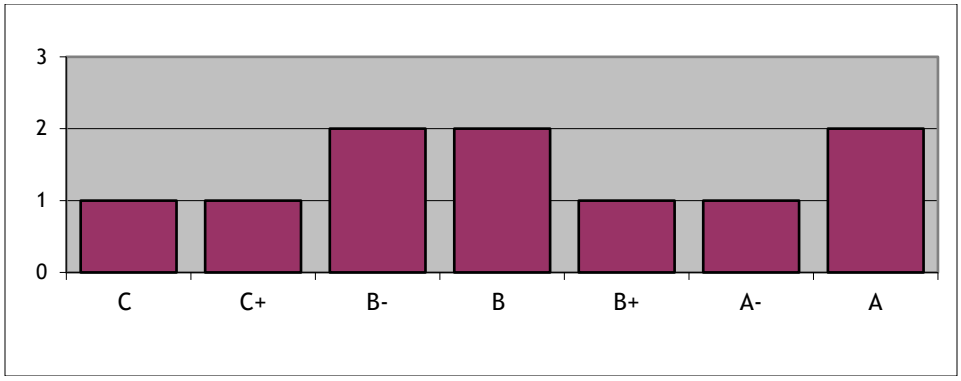


The information in the two charts is identical.

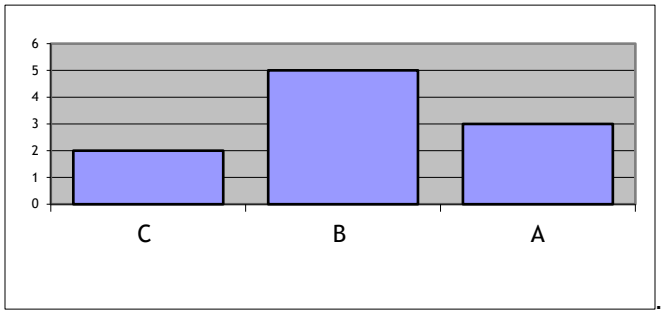
Histograms are a good way of showing how the data vary around the middle. This information about the spread of outcomes around the center is very important to most human decisions – we usually don't like risk.

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):

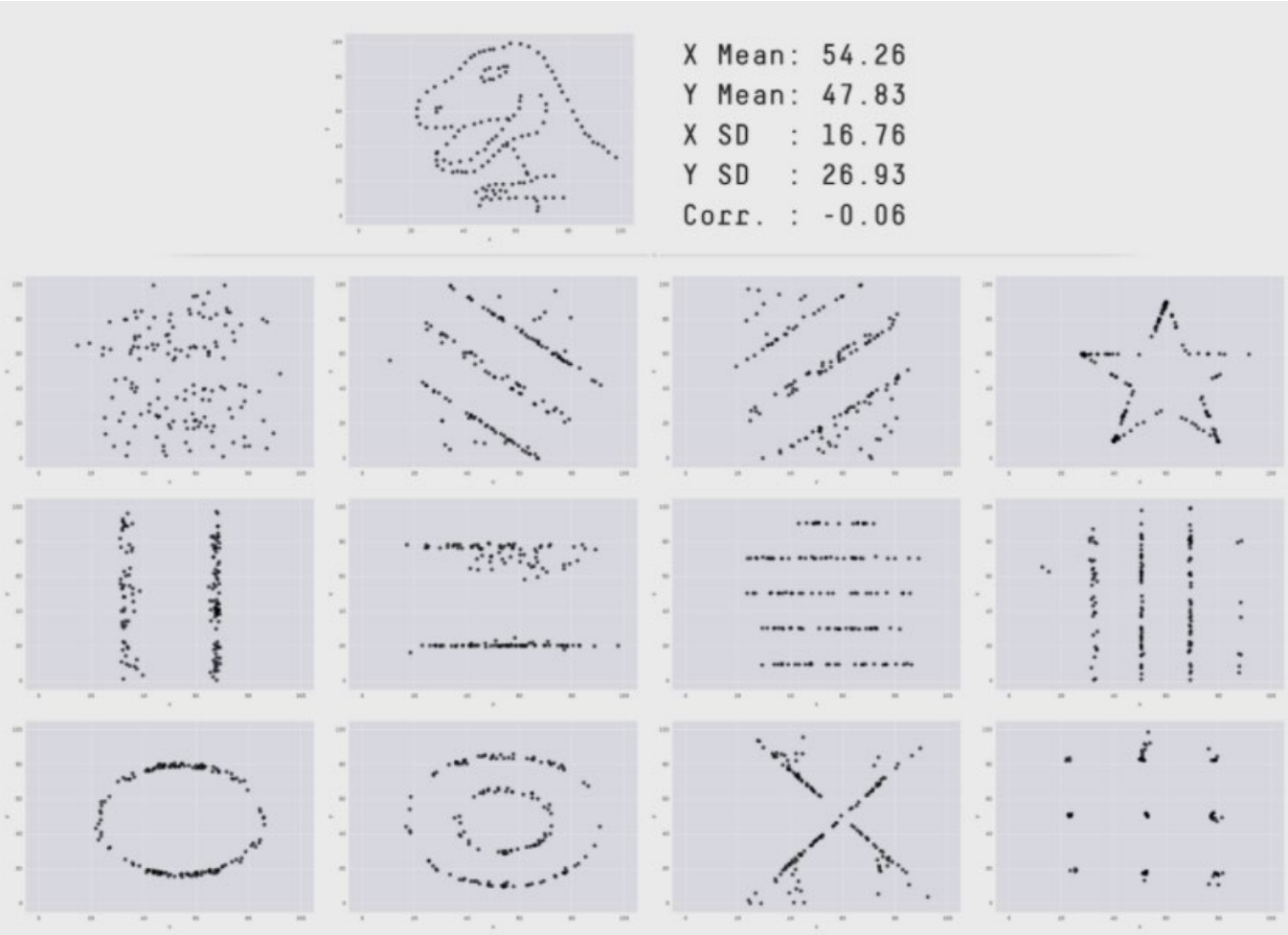


*Fig 2. The **Datasaurus Dozen**. While different in appearance, each dataset has the same summary statistics (mean, standard*

Hawkes

For much of the basics of stats, the Hawkes online material should be very useful as well.

Basic Concepts: Find the Center of the Data

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

Average

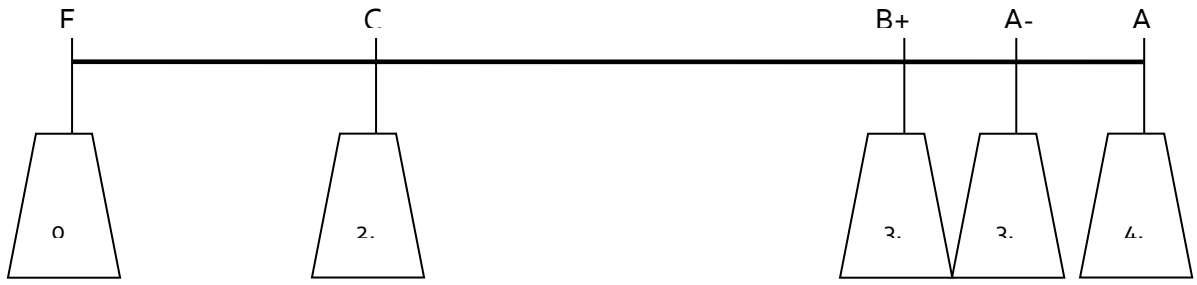
There are a few basic calculations that we start with. You need to be able to calculate an average, sometimes called the mean.

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

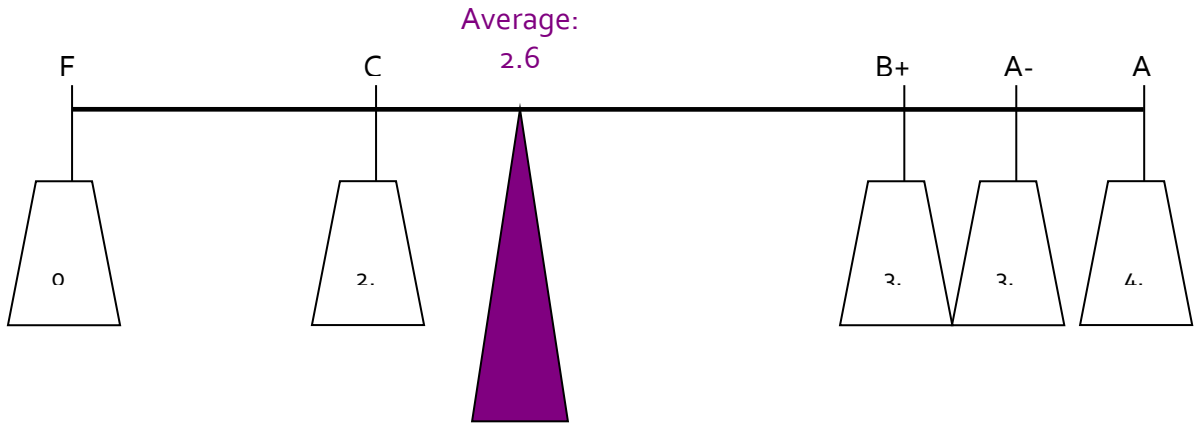
$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \, .$$

The average value of a sample is NOT NECESSARILY REPRESENTATIVE of what actually happens. There are many jokes about the average statistician who has 2.3 kids or who has a higher-than-average number of arms. 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.

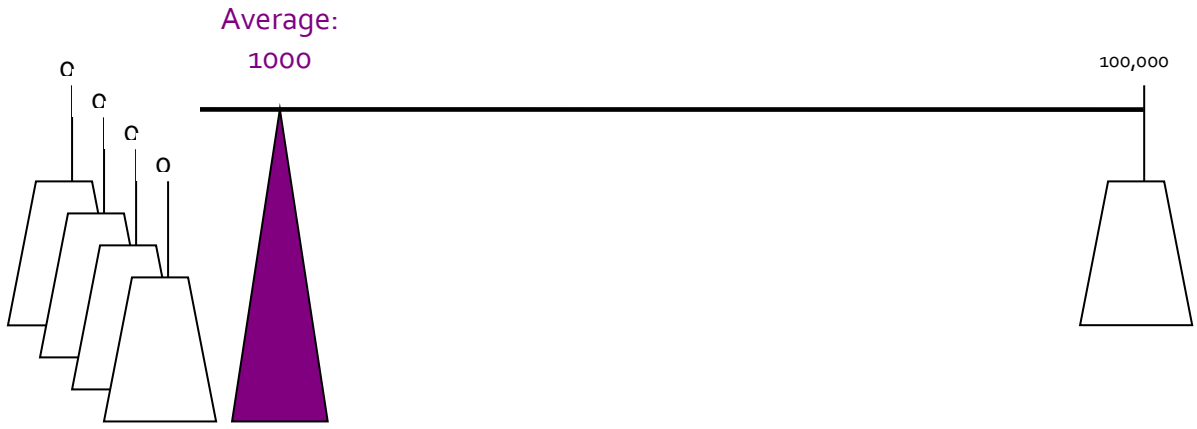
A common graphical interpretation of an average value is to interpret the values as lengths along which weights are hung on a see-saw. The average value is where a fulcrum would just balance the weights. Suppose a student is calculating her GPA. She has an A (worth 4.0), an A- (3.67), a B+ (3.33), a 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.

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

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

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

Standard Deviation

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

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

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

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

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

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

So we square the deviations and get

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

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

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

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

Of course you're asking why we bother to square all of the parts inside the summation, if we're only going to take the square root afterwards. It's worthwhile to understand the rationale since similar questions will re-occur. The point of the squared errors is that they don't cancel out. The variance can be thought of as the average size of the squared distances from the mean. Then the square root makes this into sensible units.

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

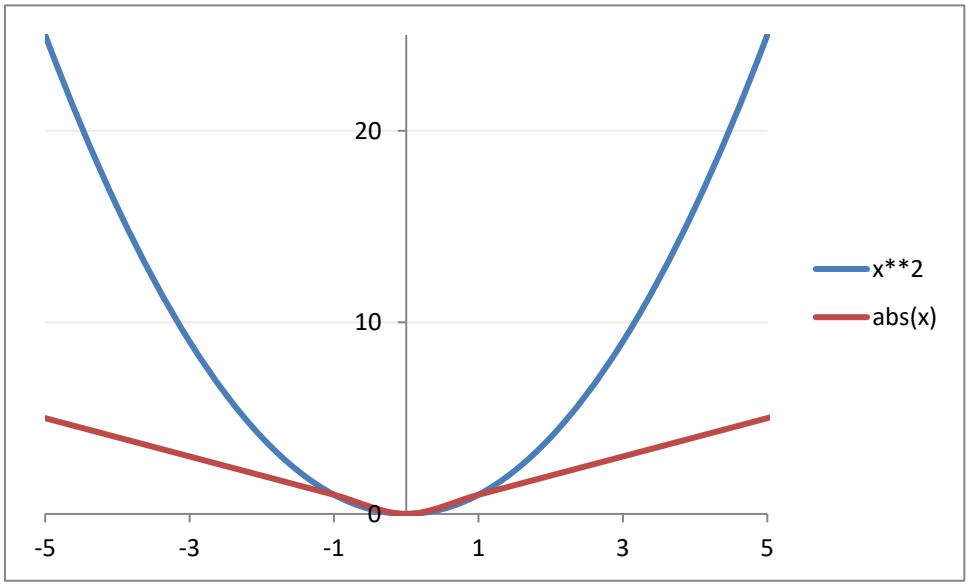
Our notation generally uses Greek letters to denote population values and English letters for sample values, so we have

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

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

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

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



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

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

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

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

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

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

Now Do It!

On Correlations: Finding Relationships between Two Variables

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

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

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

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

Run this on R …

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

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

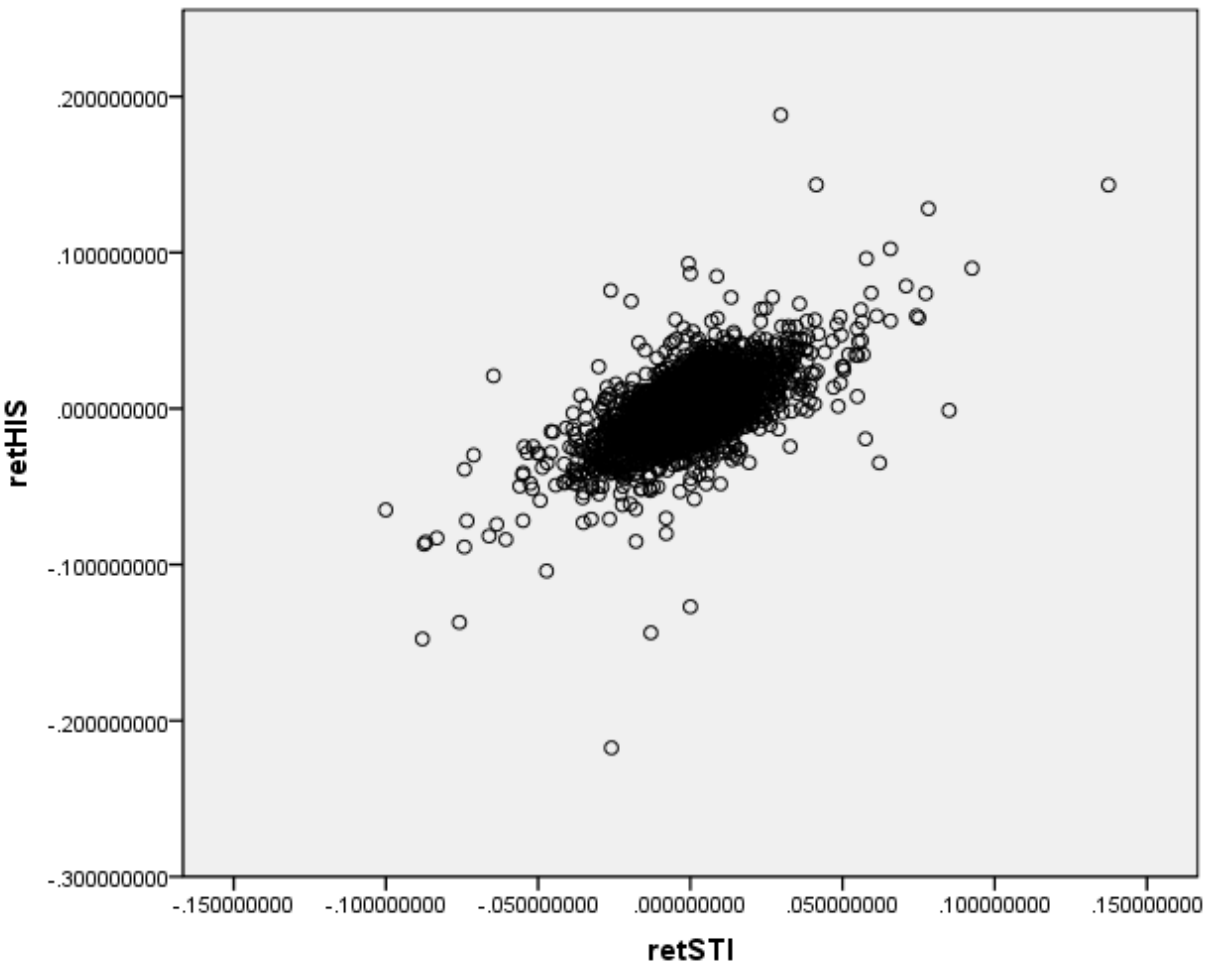
$$\overline{XY}=\frac{1}{N}\sum_{i=1}^{N}X_iY_i=\frac{1}{N}\sum_{i=1}^{N}X_i\{Y=1\}+\frac{1}{N}\sum_{i=1}^{N}X_i\{Y=0\}=\frac{1}{N}\sum_{i=1}^{N}X_i\{Y=1\}=\bar{X}_{Y=1}.$$

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

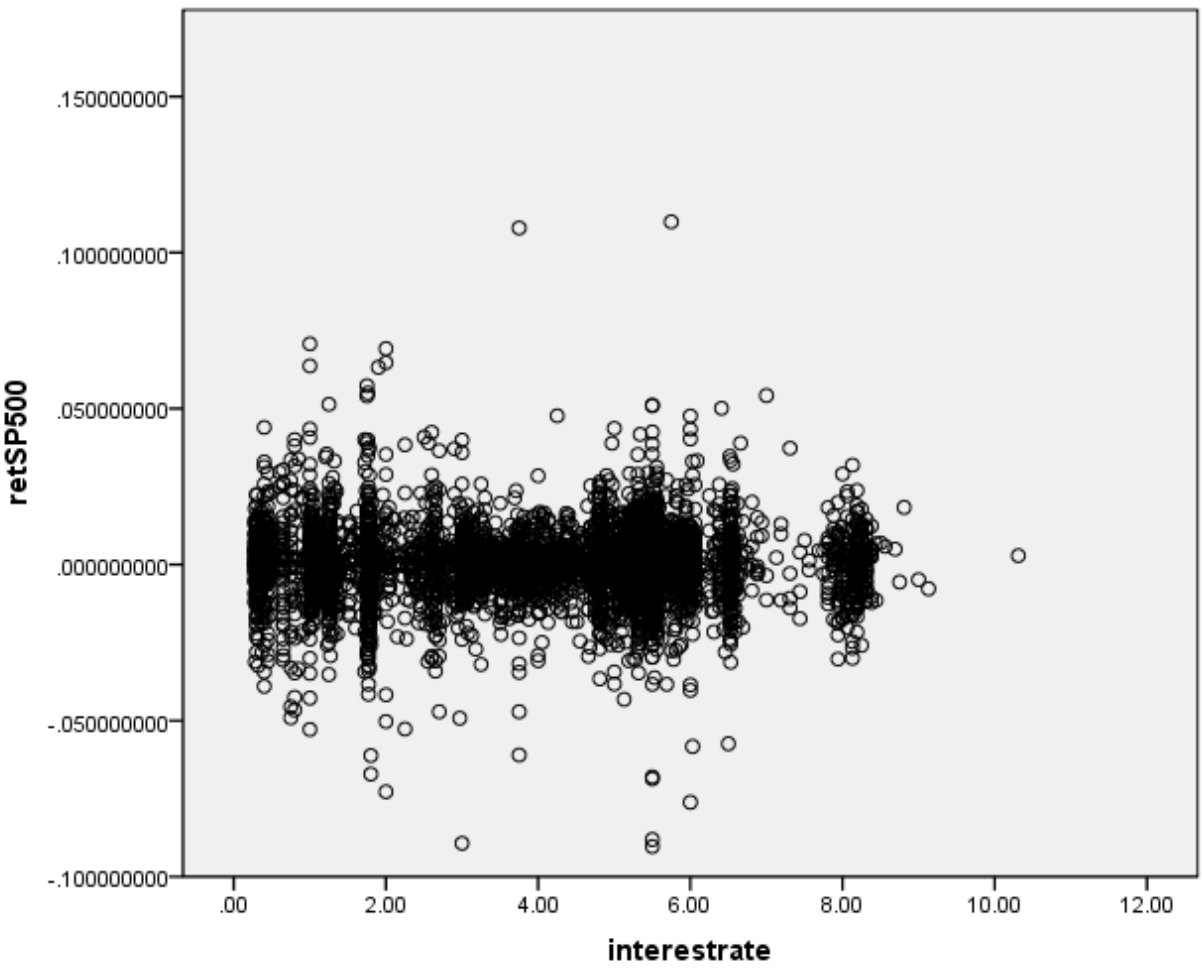
Use Your Eyes

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

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



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



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

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

How can we measure the relationship?

Correlation measures how/if two variables move together.

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

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

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

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

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

A bit of math (extra):

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

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

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

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

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

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

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

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

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

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

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

Sample covariances and sample correlations

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

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

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

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

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

X_2	σ_{21}	σ_{22}		
X_3	σ_{31}	σ_{32}	σ_{33}	
X_4	σ_{41}	σ_{42}	σ_{34}	σ_{44}

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

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

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

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.