Statistics

- We've spent the semester summarizing our data visually. This is the place to start. Humans have the ability to see visuals and understand a lot. A picture is worth a thousand summary stats.
- But often we want something more precise and concrete. We've talked about some summary stats. But we're going to be more focussed from here on out.
- We want to understand central tendency and variability. And today we're going to discuss a few options.

Measures of Central Tendency: Mean, Median, Mode.

- Imagine you're on a tennis court and you're thinking about where to stand when you're playing me. Lets say you know the distribution of my shots.
- Mode is nice in many ways. But it's most basic problem is that it's not good at describing an entire distribution. It's good at saying what's most likely, but it doesn't respond to other parts of the distribution.
- Median is also nice is so many ways. In fact, it's possible to do a lot of statistics with Median. It doesn't respond as dramatically to outliers as does Mean, among other nice properties. But it's more difficult to compute than Mean, so it's used nearly as much.
- Mean is the core measure of central tendency in statistics because it has lots of nice properties. It's simple to compute, is "smooth" in a mathematical sense, and lends itself nicely to computing variability.

Example of the Three

Measures of Variability

- Not only might we care about the center point of the data, but also the amount of movement there is in the data.
- Range (max minus min) is one way we've talked about variability in the data. You look at how far apart the biggest and smallest thing are. This is nice because it's simple but it's also not nice mathematically and it also ignores a lot in the data.
- Instead we prefer to use other measures of variability, which are centered on the center point. You can use median as a center point, but like we mentioned before, it's not so good mathematically, so we almost never use it for variability. There's a branch that does some stuff with it. But we're going to center on the mean.
- Imagine you're on a tennis court and you're thinking about how much running you'll have to do when you play me. Lets say you know the distribution of my shots and stand in the middle, at the mean left-right location on the court.
- How far will you have to run on average? Well we could just take the difference between where you're standing and where you would run to then take the average of this difference.
- Again we're left with the question of which average to use. Some statisticians will sometimes use median. But for so many reasons, this becomes very difficult to implement. So we nearly never use it.

- Instead, lets take the mean, giving us something we might call "mean difference". But using our example, the mean difference is exactly zero! Why?
- To fix this problem of negatives, we could take the mean absolute difference. This is a little better. But again it's not nice mathematically because of the absolute value.
- Instead what we've agreed on across fields is to use the square: variance = $\sum (x_i - \bar{x})^2 = (X - \bar{X})^2 / n$. This fixes the absolute value problem, since the square of a negative is positive. And it is nice to work with mathematically.
- The problem with THIS though, is that the units are strange. . Instead, we use the square root of this thing, called standard deviation. $\sigma = \sqrt(\sum (x_i - \bar{x})^2) = \sqrt((X - \bar{X})^2 / n)$.
- So if we know the distribution of our data, then we would be able to compute mathematically the measures of central tendency and variability of the data without having to do any work.
- Show some example distributions and measures of central tendency.
- This makes it easy to answer so many questions about the data.
- Random variables here.

4. this is what we call a random varible: we draw an observation from a distribution

Sampling Outline

- In some cases, when the distribution or our question is very wonky, we actually cannot solve these things mathematically. Nate Silver, for example, who built FiveThirtyEight, uses a very complex statistical model of elections that's too difficult to solve mathematically. So in these cases, we run simulations. We make thousands of draws from a distribution and compute the measures we care about from this data, realizations of a random variable.
- The random variable captures every possible way the world could go. The realization of the random variable is ONE PARTICULAR way the world went.
- You may feel like the simulation approach somehow breaks a rule in statistics. But it doesn't. Let me show you that this is valid by simulating the distributions of some known distributions.
- So in cases when we have very complex distributions, this computer simulation approach is very nice and allows us to build more accurate models (meaning complex beyond what's possible mathematically).
-

1. we will only ever see a sample of a population. even with the census, we're taking a sample - no way to know every possible census we could have taken, there's still variability in the way the world could go
2. the sample is a subset (repeated draws) of the population
3. the population follows a distribution
4. a sample is just repeated realizations of a random variable

5. as we talked about before, if we know the distribution, our life is simple. we can measure whatever we like. how many people have ages above 50? what's the probability a randomly drawn person will be younger than 30? what's the probability a randomly drawn person will have an extreme age (under 10, above 90)? what is the average age in the population? we can simulate an answer by repeated sampling or we can use the mathematical properties of the distribution to integrate under the curve. and we could do this for any distribution we'd like.
6. but in practice we NEVER know the distribution

8. so what do we do? how would we answer the question: is the average age in pittsburgh greater than 45? can we answer it without asking everyone and without knowing the true distribution of ages?
9. yes. lets run a simulation. i have an unknown distribution. lets take repeated samples.
10. we can look at the distribution of one sample, show the mean, and compare it to 45.
11. what happens if we do this a bunch of times? well sometimes we get True and sometimes we get False. so which is it?
12. well, how comfortable are we being wrong? like, with the question at hand and the data, it looks like sometimes my guess is correct and sometimes its not.
13. we can be less wrong if we increase the sample size - if i do repeated samples, i can see that i'm less wrong - so if I draw once, i'm going to be wrong with a lower probability
14. so the distribution of sample means has some known properties.
15. it turns out we know exactly THIS distribution - do a bit of derivation of the CLT
16. then show that the CLT estimate gives the same conclusion as doing a bunch of sampling
17. it's mathematically identical to move the confidence interval from the truth to the sample mean, and show that we draw the same conclusions
18. so we don't even need to know anything about the underlying population? thats right! this is the most unhearalded idea in modern science. we can say so much while starting with so little.
19. then do a t-test using regression, connecting the p-value to the number of times we got the wrong answer
20. there are lots of other ways to do (or explain) this, but they all mean the same thing: the central limit theorem, the sample mean follows a normal distribution with some standard deviation. but in practice we don't know the standard deviation. using the standard deviation of the sample introduces additional variability, which is captured in the t-distribution.