PS599 - Introduction to Empirical Methods - Section Notes (Extremely Abridged!)

Jason S. Davis

Fall 2016

Contents

1 Welcome/Introduction

- Jason Davis - E-mail: jasonsd@umich.edu - Office: Haven Hall 7730 - Telephone: (write in class)
- Need to determine office hours. I'm thinking Thursdays 2-4PM?
- Donuts last section from Dimo's Deli and Donuts!
- Problem Set 1: I added more proofs to give you more practice, because they come up a bunch in 699. Indeed, some of the proofs you're doing will probably come up again in the first 699 problem set! It's pass/fail, and I don't expect to use the fail option very much (if at all) so you can consider this a low stakes opportunity to get some practice and some feedback. Like math camp!

2 Set Theory

- Set theory is prior to probability theory.
- Without getting too deep into the details, in set theory we can talk about "sigma algebras" on a set X, which are a collection of subsets Σ that satisfy the following properties.
  1. $\varnothing \in \Sigma$ (contains the empty set)
  2. $A^c \in \Sigma$ if $A \in \Sigma$ (closed under complement)
  3. If A_i $i \in \{1, \ldots, n\}$ are in Σ, then $\bigcup_{i=1}^{n} A_i$ is in Σ (closed under union)
- Consider the set $X = \{1, 2, 3, 4\}$.
- The smallest sigma algebra would be $\Sigma = \{(\varnothing), (1, 2, 3, 4)\}$, since $\varnothing^c = X$
- The largest would be the power set $P(X)$, which is the set of all subsets.
- Are the following sigma algebras?
  1. $T = \{\varnothing, (1, 2, 3, 4), (1)\}$? Nope! Not closed under complement.
  2. $U = \{\varnothing, (1, 2, 3, 4), (1), (2, 3, 4)\}$ Yep! Satisfies all properties.
  3. $U = \{\varnothing, (1, 2, 3, 4), (1), (2, 3, 4), (2), (1, 3, 4)\}$? No, not closed under union: $(1) \cup (2) = (1, 2)$, which is not in the collection of subsets.
- Measure theory, which comes up sometimes in 699 but you don't really need to know thoroughly, is about assigning "measures" to these collections of subsets. You need a σ-algebra to assign measures (a set without a σ-algebra is non-measurable).
- Probabilities are just a particular kind of measure. In fact, they have the same axioms as any old measure, but with one added: the fact that probabilities are bounded above by 1 (we'll revisit this when we take a look again at the axioms).
- You can kind of see how the sigma algebra fits in to thinking about probability. The subsets in the algebra become the "events" over which probabilites are defined.
- **Partition:** A collection of nonempty subsets such that each element in the set is contained in one and only one subset.
- Note: this implies (1) the subsets are disjoint; (2) they "cover" the set, in some sense.
- Example: For number line $[0, 1]$, a partition would be $\{[0, 0.5), [0.5, 1]\}$.
- $\{[0, 0.5], (0.5, 1]\}$ would not be a partition, because the sets are not disjoint.

- $\{[0, 0.5), (0.5, 1]\}$ would not be a partition, because 0.5 is not in any subset.
- Some properties of sets are as follows:
  - $A \cup B = B \cup A$
  - $A \cap B = B \cap A$
  - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  - $(A \cup B)^c = A^c \cap B^c$
  - $(A \cap B)^c = A^c \cup B^c$
  - For any set B, $A = (A \cap B) \cup (A \cap B^c)$
  - $A \cup B = A \cup (B \cap A^c)$
- Keep in mind that you can assume all of these properties for the proofs in the problem set! What you can't assume are properties of probabilities, e.g. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This is the kind of thing you need to prove.

3 Probability Theory & Proofs

- Probability axioms:
  1. For any event A, $P(A) \geq 0$
  2. If Ω is the sample space, $P(\Omega) = 1$ (this is the one that makes it different from other measures)
  3. For a series of disjoint events A_i with $i \in \{1, \ldots, n\}$, $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$
- When people write things like "You have a probability space "$(\Omega, \mathbb{F}, \mathbb{P})$", the Ω is the set it's defined on (sample space), the $\mathbb{F}$ is the sigma algebra/sigma field (i.e. the set of events on which probabilities are defined), and $\mathbb{P}$ is the probability measure defined on those subsets. So that's the quick and dirty measure theory behind the language you'll see sometimes in 699.
- What is a proof? How do you write one?
- In general, you are given a set of things that you assume, and then want to derive, using the logical rules you learned in math camp, a particular conclusion from those assumptions.
- "Direct" proofs, generally, just show one step leading to the next one. $LS = RS$ type stuff is an example of a direct proof.
- More complicated proof strategies include proofs by contradiction, proofs by induction, etc. We will not do that here, though if you are doing the game theory sequence, we will probably get to it in 681 at least.
- For now, the biggest complications for the problem sets are (1) what can you assume?; (2) how do you go about actually constructing the proof.
- As mentioned earlier, you can assume all the stuff from set theory. So what you want to do is use those rules, in combination with the things in the probability axioms, to generate the conclusions you want.
- Once you've proven one property, you can use it to prove others.
- So, let's do some examples:

1. $P(A^c) = 1 - P(A)$
   Ans: By definition, $\Omega = A \cup A^c$. Since A and A^c are disjoint, $P(A \cup A^c) = P(A) + P(A^c) = P(\Omega) = 1$. Rearrange to get above.
2. $P(\varnothing) = 0$.
   Ans: Using (1), we have that $P(\varnothing) + P(\varnothing^c) = 1$. Rearrange to get $P(\varnothing) = 1 - P(\Omega) = 0$.
3. If $A \subset B$, then $P(A) \leq P(B)$
   Ans: First, note that $A = (A \cap B) \cup (A \cap B^c)$ and $B = (B \cap A) \cup (B \cap A^c)$. Now, consider that because $A \subset B$, $A \cap B^c = \varnothing$. Thus $P(A) = P(A \cap B)$ and $P(B) = P((B \cap A) \cup (B \cap A^c))$ which by the union property of probabilities (given that these are disjoint), gives us $P(B) = P(A \cap B) + P(B \cap A^c)$. Since $P(B \cap A^c) \geq 0$ by non-negativity of probabilities, it is the case that $P(B) \geq P(A)$.

4 Distribution review, variance, covariance

- What is a parameter?
- What is a statistic?
- What is an estimate?
- What is an estimator?
- Parametric families of distributions? Remember binomial $P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$
- Linear combinations of independent normal random variables produce new normal random variables.
- Basic review of what variance and covariance are.

4.1 Expected Values

1. $E(a) = a$
2. $E(bX) = bE(X)$
3. $E(a + bX) = a + bE(X)$
4. $\Sigma E(g(X)) = E(\Sigma g(x))$
5. $E(E(X)) = E(X)$

4.2 Variance and Other Moments

- m^{th} moment of X is $E(X^m)$. m^{th} central moment is $E(X - E(X))^m$
- Variance is second moment.
- Covariance: $E(x - E(x))(y - E(Y))$
- Also can be expressed as $E(X^2) - (E(X))^2$. See proof below.

$$\begin{aligned} Var(X) &= E(X - E(X))^2 \\ &= E(X^2 - 2xE(X) + (E(X))^2) \\ &= E(X^2) - E(2xE(x)) + (E(X))^2 \\ &= E(X^2) - 2E(x)E(x) + (E(X))^2 \\ &= E(X^2) - (E(x))^2 \end{aligned}$$

4.3 Rules of Variance and Covariance

- $Var(a + bX) = b^2 Var(X)$
- $Var(a + bX + cY) = b^2 Var(X) + c^2 VAR(Y) + 2bcCov(X, Y)$
- $Var(c) = 0$
- $Cov(X, Y) = E(XY) - E(X)E(Y)$. Note, this is zero if X and Y are independent, as in this case $E(XY) = E(X)E(Y)$
- $Cov(X + c, Y + b) = Cov(X, Y)$
- $Cov(cX, bY) = cbCov(X, Y)$
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
- $Cov(X, X) = Var(X)$

5 Causal Inference

- In class, Walter talks a lot about causality versus association, etc. I think it's useful to have a bit of an understanding of what people mean when they say causal inference, and a few of the faultlines in the discipline.
- Rubin Causal Model/Potential Outcomes: Would like to see what would happen, for one individual, if they received the "treatment" rather than the "control". But one of these potential outcomes is always obviated by receiving the other. This is described as the "fundamental problem of causal inference."
- Aside: I once was assigned a medication by my doctor. After I'd finished taking it, he asked "did it work?" I then tried to explain that I have no way of knowing that, as I could only observe what happened in the state of the world where I took it, where I recovered from my illness. I have no way of knowing what would have happened if I didn't take the medication. He nodded and said "that's right, n of one". But that's not really the point! We have RCTs that say the medicine works on average, but for any individual we can't really assess the treatment effect. But I digress...
- In general, causal inference people jump off from this, and say you can't really do "causal" work if you can't imagine assigning both treatment and control in this fashion.
- So for instance, they would tend to say that assessing the causal impact of "gender" or "sex" is essentially impossible, because it implies fundamentally changing the unit you're applying the treatment to.
- You might be able to assess something like having a recognizably male or female name would have a causal impact. But that's a little different.
- This doesn't mean that finding systematic differences between men and women that can't be attributed to a bunch of other factors isn't important.
- In the IR literature, people talk about the "democratic peace". It's likely causal inference people would be skeptical of the idea that you can really imagine tranforming a country from a nondemocracy to a nondemocracy in the 'treatment/control" sense. But the associations are still important.
- Moreover, in game theory, you often get predicted associations between variables that come out of the model, called "comparative statics". This "generically violates" causal inference, as Walter said, because it is a product of units that are interacting, not something where you could just isolate the effect of x on y.

- When Walter asked about whether experiments should be the "gold standard" of research, I interpret his question as not so much being about practical problems of experiments (e.g. external validity, response rates, expense, whatnot) but about whether or not this "treatment vs control" approach is really the way to think about social science.
- Some people definitely think yes! And there are lots of neat insights that come out of the causal inference literature. But if you're dealing with situations where you have a lot of spillovers/diffusion, strategic interactions, etc. then you may be more interested in thinking about things differently.
- Partial equilibrium versus general equilibrium.
- SUTVA - stable unit treatment value assumption. Implies that one unit should be unaffected by treatment assignment to other units.
- Violated when you have spillovers, etc.

6 Transformations of Random Variables

- In some cases, we may be interested in the distribution of some function of one or more random variables.
- For instance, consider that a model is taking a bunch of random variables, and then via some function, producing a new one. That new one will have a distribution of its own.
- Example: Say we run a shawarma shop, and our profits are a function of the cost of vegetables C, and the number of sandwiches bought Q, e.g. $\pi = 5Q - CQ$, and both C and Q are random variables with some distribution (exchange rates affect price of vegetables in Canada!). We might want to know the distribution of our profits.
- Rules from before: linear combinations of normal random variables will also be normal. *This is true whether or not they are independent.* Multivariate normal specifies correlations between normal variables in a variance-covariance matrix. Something to talk about later.
- If they are independent, can just use the variance and covariance rules from before.
- E.g. from above, imagine if $Q \sim N(5, 3)$ and $C \sim N(2, 0.5)$. What would be our profit distribution?
- However, this isn't true for all random variables; indeed, Walter referred to it as "magic".
- However, to demonstrate that this is not always sufficient, consider a variable $D \sim Poisson(4)$. Now say we take a transformation $W = 10D$. Is this going to be Poisson?
- $E(5D) = 10E(D) = 10(4) = 40$. $Var(5D) = 5^2 Var(D) = 25(4) = 100$. Since for Poisson, $\lambda = E(X) = Var(X)$, this can't be Poisson.
- For transformations of discrete random variables, you are also generally changing the values on which it is defined.
- Consider a random variable X that is distributed discrete uniform between 1 and 6 (i.e. a die). What would be the distribution of $Y = a + X$?
- $Pr(X = x) = 1/6$ for $x \in \{1, \ldots, 6\}$. $Pr(Y = y) = 1/6$ for $y \in \{a + 1, \ldots, a + 6\}$
- Say $X \sim Binom(10, 0.5)$. What's the distribution of $Y \sim 5X$?
- Binomial gives $Pr(X = k) = \binom{10}{k}(0.5)^k(0.5)^{10-k} = \binom{10}{k}(0.5)^{10}$

- Want to find $Pr(Y = c)$, where c will now be defined on $c \in \{0, 5, 10, \ldots, 50\}$ because of the transformation.
- $Pr(5X = c) = Pr(X = \frac{c}{5})$.
- So we get $Pr(Y = c) = \binom{10}{c/5}(0.5)^1 0$ for $c \in \{0, 5, 10, \ldots, 50\}$
- Example: $Pr(Y = 20) = \binom{10}{20/5}(0.5)^1 0 = \binom{10}{4}(0.5)^1 0$, which you'll notice is the same probability as the $Pr(X = 4)$. Which is what we want, because the function $(Y = 5X)$ maps $X = 4$ to $Y = 20$. So the probabilities SHOULD be the same.
- More generically, for some function of a discrete random variable $Y = g(X)$, $Pr(Y = y) = Pr(g(X) = y) = Pr(X = g^{-1}(y))$, where g^{-1} is the inverse function of g, if we assume g is a bijection, i.e. is one-to-one. Why would this matter?
- If g is not a bijection, things get a little more complicated, as you have to sum over all the values that map back.
- Back to the die example: imagine your function is $Y = g(X) = (X - 3)^2$. Now the values the function can take on are $\{0, 1, 4, 9\}$. $Pr(0) = Pr(9) = 1/6$. $Pr(1) = Pr(4) = 1/3$. For 1 and 4, had to sum over the probabilities for the different values of X that map to values of Y.
- There are different, analogous techniques when you have a continuous distribution. And they get more complicated when you do transformations that involve multiple random variables. Won't get too deep into it.
- Say for instance, say you have a random variable with CDF $F(X) = \frac{1}{4}x^2$ for $x \in (0, 2)$, and thus a pdf of $f(x) = \frac{1}{2}x$.
- If we wanted to find the pdf of $Y = X^2$? Consider $F(Y \leq y) = F(X^2 \leq y) = F(X \leq \sqrt{Y})$.
- Plugging it to the CDF, we would get $F(Y) = \frac{1}{4}(\sqrt{Y})^2 = \frac{1}{4}Y$. We can also see that the domain of Y changes to $(0, 4)$.
- Then, to find the PDF, just take the derivative. So $F'(Y) = f(Y) = \frac{1}{4}$. Which is what kind of distribution? (Ans: Uniform between 0 and 4)
- This works if your function is monotonic (as $F(X) = X^2$ is). If it isn't, you'll need to partition up the domain, to do the analogous thing to summing that you did with the discrete variable. Also need a continuous derivative.
- Proving that a squared standard normal distribution is distributed chi-squared is an example of this. The domain needs to be split at zero for X^2 to be monotonic on each part. Below zero: more negative lead to higher X^2. After zero: more positive leads to higher X^2.
- Consider this: each positive value's probability gets doubled in some sense, and the domain changes so there are no negative values.
- Won't get into this in detail, but it's good for you to have some intuition.

7 Distributions

- Walter said a bunch of stuff I would have otherwise gone over.
- Poisson used a lot for count data.

- These might be important if you're thinking of some kind of data that can be directly modeled using one of them. So things like "numbers of phone calls per hour" have been modeled using Poisson distributions.
- However, many social science questions are more complicated than just trying to describe how something is distributed.
- These distributions become important for more complicated questions when we think about "wrapping them" around a model.
- So for instance, if you have a linear model that looks like $Y = \beta_0 + \beta_1 + \beta_2 + \epsilon$ but you want to bound the values between zero and one, you might wrap a normal distribution around it. Normal distribution has as its domain all real numbers.
- These are logits and probits! And they're the kind of thing you estimate using MLE. So you can start to see how this useful.

8 Estimation

8.1 Notes on Problem Set 3

- Problem set should have both an analytical and empirical component for each estimation method.
- In the analytics, you show how you find the *estimator*, using the estimation method discussed. In each case, this should give you a relatively simple expression/statistic, involving for instance taking the sum of the observations and dividing by the number of observations.
- Many of the estimators produced by the different estimation procedures will be the same!
- The empirical component involves using R to compute the *estimates* for each parameter from the data given to you. This should be a very, very simple exercise, as the expressions you get will be very easy. If you find yourself using optim, nlm, or rgenoud, you've gone too far!
- To put this in perspective: you could do all the calculations on a hand calculator if you wanted to.
- Walter's "normalMLE.R" file does this in the first part, then moves on to using R to do symbolic math, or using R to compute MLE estimates using numerical algorithms.
- If you want to learn how to find MLEs numerically go ahead, but that is not what we are doing for this problem set!

8.2 Maximum Likelihood Estimators

Poisson Distribution

The Poisson Distribution is:

$$= \frac{e^{-\lambda}\lambda^{x}}{x!}$$

So we construct the likelihood function:

$$L(\lambda|\boldsymbol{x}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

To maximize this likelihood, it will be easier to work with the log-likelihood function (use base e log):

$$\begin{aligned} logL(\lambda|\boldsymbol{x}) &= \Sigma_{i=1}^{n} log\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \\ &= \Sigma log(e^{-\lambda}) + \Sigma log(\lambda^{x_i}) - \Sigma log(x_i!) \\ &= \Sigma(-\lambda) log(e) + \Sigma x_i log(\lambda) - \Sigma log(x_i!) \\ &= -n\lambda + \Sigma x_i log(\lambda) - \Sigma log(x_i!) \end{aligned}$$

Taking the derivative with respect to λ:

$$\frac{\partial logL}{\partial \lambda} = -n + \frac{\Sigma x_i}{\lambda} + 0$$

Set to zero and rearrange to get:

$$\lambda = \frac{\Sigma x_i}{n}$$

Which is the maximum likelihood estimator of λ in the Poisson distribution.

Binomial Distribution with $n = 15$

We're given $n = 15$ in the problem set, so the only unknown parameter is p. Thus we have:

$$= \binom{15}{x} p^x (1-p)^{15-x}$$

Resulting in the likelihood function (where m is the number of observations)

$$L(p|\boldsymbol{x}, n = 15) = \prod_{i=1}^{m} \binom{15}{x_i} p^{x_i} (1-p)^{15-x_i}$$

And the log-likelihood function (all Σs are $\Sigma_{i=1}^{m}$):

$$logL(p|\boldsymbol{x}, n = 15) = \Sigma x_i log(p) + \Sigma(15 - x_i) log(1-p) + \Sigma log\left(\binom{15}{x_i}\right)$$

Taking the derivative and setting to zero

$$\begin{aligned} \frac{\partial logL}{\partial p} &= \frac{\Sigma x_i}{p} + \frac{\Sigma 15 - x_i}{1-p}(-1) = 0 \\ \leftrightarrow \frac{1-p}{p} &= \frac{\Sigma(15 - x_i)}{\Sigma x_i} \\ \leftrightarrow \frac{1}{p} - 1 &= \frac{\Sigma(15 - x_i)}{\Sigma x_i} \\ \leftrightarrow \frac{1}{p} &= \frac{\Sigma(15 - x_i) + \Sigma x_i}{\Sigma x_i} \\ \leftrightarrow p &= \frac{\Sigma x_i}{\Sigma(15) - \Sigma x_i + \Sigma x_i} \\ \leftrightarrow p &= \frac{\Sigma x_i}{m(15)} \end{aligned}$$

8.3 Method of Moments

Poisson Distribution
Method of moments estimation involves comparing sample moments, where the jth sample moment is straightforwardly calculated as $\frac{\Sigma x_i^j}{n}$, to the theoretical moments. Note that the theoretical moment will be computed in terms of the parameter(s), so we can use this approach to estimate these paramters. We use as many moments as there are parameters to ensure that we have a system of m equations with m unknowns, allowing us to solve for each parameter. With the Poisson distribution, there is only one parameter, so we only need to look at the first moment.

$$E(x) = \int_{-\infty}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} dx$$

We can look up the theoretical moment rather than computing it ourselves. Substituting in the formula for the theoretical moment and the formula for the sample moment, we get:

$$\frac{\Sigma x_i}{n} = \lambda$$

Which, incidentally, is the same as what we obtain by MLE.

Binomial Distribution
We are given $n = 15$, and given that the general form for $E(x)$ is $E(x) = np$ when x is a binomially distributed random variable, we have the theoretical moment $E(x) = 15p$. Comparing to the sample moment, we get (with m as the length of the sample vector):

$$\frac{\Sigma x_i}{m} = 15p$$
$$\leftrightarrow p = \frac{\Sigma x_i}{15m}$$

Which it should be noted is equivalent to the estimator derived by MLE.

Binomial Distribution with unknown n

In this section, I am going to do something that is technically wrong, but still illustrates the procedure you need. Namely, I'm going to pretend that the N of the binomial distribution can vary continuously. Obviously you can't have fractional numbers of trials; it has to be an integer. However, if you were to apply this approach to something like, say, the normal distribution, where neither parameter has to be an integer, it would work out fine.

Here, we need to estimate two parameters, n and p. Given this, we need to find two equations so we can solve for the two unknowns, and using method of moments, these equations will be (1): $E(x^1) = \frac{\Sigma x_i}{n}$ and (2): $E(x^2) = \frac{\Sigma x_i^2}{n}$ (these equations set the sample moments equal to the theoretically-derived moments). By looking it up in a book or on google, we can find our theoretically-derived moments, namely that $E(x^1) = np$ and $E(x^2) = np(1 - p + np)$, and substitute in. Thus our two equations are:

$$\frac{\Sigma x_i}{m} = np \leftrightarrow \hat{p} = \frac{\Sigma x_i}{mn} \quad (1)$$
$$\frac{\Sigma x_i^2}{m} = np(1 - p + np) \quad (2)$$

So let's substitute the expression for p into equation two and do some tedious algebra.

$$\begin{aligned}
\frac{\Sigma x_i^2}{m} &= n\frac{\Sigma x_i}{mn}\left(1 - \frac{\Sigma x_i}{mn} + n\frac{\Sigma x_i}{mn}\right) \\
&= \frac{x_i}{m}\left(1 - \frac{\Sigma x_i}{mn} + \frac{\Sigma x_i}{m}\right) \\
\leftrightarrow \frac{\frac{\Sigma x_i^2}{m}}{\frac{\Sigma x_i}{m}} &= \frac{\Sigma x_i^2}{\Sigma x_i} = 1 - \frac{\Sigma x_i}{mn} + \frac{\Sigma x_i}{m} \\
\leftrightarrow \frac{\Sigma x_i}{mn} &= 1 - \frac{\Sigma x_i^2}{\Sigma x_i} + \frac{\Sigma x_i}{m} \\
\leftrightarrow \hat{n} &= \frac{\Sigma x_i}{m\left(1 - \frac{\Sigma x_i^2}{\Sigma x_i} + \frac{\Sigma x_i}{m}\right)} \\
&= \frac{\Sigma x_i}{m - m\frac{\Sigma x_i^2}{\Sigma x_i} + \Sigma x_i}
\end{aligned}$$

So now we can take our estimators to the data. Let's first compute our estimate of n, then we can compute our estimate for p straightforwardly from $\hat{p} = \frac{\Sigma x_i}{m\hat{n}}$ Here's some code I wrote to generate these estimates.

```
data<-read.csv("pset3.csv")
sigx<-sum(data$xBinom)
sigx2<-sum(data$xBinom^2)
m<-length(data$xBinom)
n<- sigx/(m - m*(sigx2/sigx)+ sigx)
p <- sigx/(m*n)
```

This gives estimates of 14.2 for $\hat{n}$ and 0.7038 for $\hat{p}$. These are the point estimates we get when we don't know the value for n. And, of course, the estimate for n doesn't make sense because it's not an integer, and we'd have to do more complicated stuff to make it actually work. In any event, Walter gives you n which means you only have to find $\hat{p}$, and you will get a different (slightly, you get 0.6688 to be precise) value for p than you do when you don't know a priori what the value of n is and are instead just fitting the best binomial distribution you can to the data. However, one of the distributions you are asked to estimate the parameters of has two parameters, so you will need to adopt this kind of approach.

8.4 Least Squares Estimation

For the mean, generally:

$$f(\mu) = \Sigma_{i=1}^{n}(x_i - \mu)^2$$

Take derivative and set to zero

$$\begin{aligned}
\frac{\partial f}{\partial \mu} &= -2\Sigma(x_i - \mu) = 0 \\
&= \Sigma x_i - \Sigma\mu \\
\leftrightarrow \Sigma x_i &= n\mu \\
\leftrightarrow \mu &= \frac{\Sigma x_i}{n}
\end{aligned}$$

Poisson Distribution
We have a formula for the theoretical expected value (mean) and can thus substitute that in and solve for the parameter value desired.

$$f(\lambda) = \Sigma_{i=1}^{n}(x_i - \lambda)^2$$

Take derivative and set to zero

$$\begin{aligned}\frac{\partial f}{\partial \lambda} &= -2\Sigma(x_i - \lambda) = 0 \\ &= \Sigma x_i - \Sigma\lambda \\ \leftrightarrow \Sigma x_i &= n\lambda \\ \leftrightarrow \lambda &= \frac{\Sigma x_i}{n}\end{aligned}$$

Binomial Distribution ($n = 15$)
We have a formula for the theoretical expected value (mean) and can thus substitute that in and solve for the parameter value desired.

$$f(\lambda) = \Sigma_{i=1}^{n}(x_i - 15p)^2$$

Take derivative and set to zero

$$\begin{aligned}\frac{\partial f}{\partial p} &= -2(15)\Sigma(x_i - 15p) = 0 \\ &= \Sigma x_i - \Sigma 15p \\ \leftrightarrow \Sigma x_i &= (15)(np) \\ \leftrightarrow p &= \frac{\Sigma x_i}{15n}\end{aligned}$$

Normal distribution estimate for $\boldsymbol{\sigma^2}$
This is slightly different; σ^2 is the average squared deviation, so we need to compare this to the sample squared residual, using our estimate of μ.

$$f(\sigma^2) = \Sigma_{i=1}^{n}((x_i - \mu)^2 - \sigma^2)^2$$

Take derivative and set to zero

$$\begin{aligned}\frac{\partial f}{\partial \sigma^2} &= -2\Sigma((x_i - \mu)^2 - \sigma^2) = 0 \\ \leftrightarrow \Sigma(x_i - \mu)^2 &= \Sigma\sigma^2 \\ \leftrightarrow \sigma^2 &= \frac{\Sigma(x_i - \mu)^2}{n}\end{aligned}$$

8.5 Properties of Estimators

- MLE, method of moments, least squares, etc. all gives us estimators for parameters.
- Are they good estimators? What would that mean?
- A topic that we don't get into in too much depth in this course is properties of estimators.
- These include:
  - Unbiasedness: expectation equal to the parameter.
  - Consistency: estimators converges in probability to parameter as n goes to infinity.
  - Efficiency: amongst unbiased estimators, has lowest variance.
  - Robustness: can mean a number of things. Usually has to deal with whether it works for broader range of assumptions.

- Sometimes you'll see things like corrections, e.g. for sample variance dividing by $n-1$ instead of n. These are to deal with bias.
- However, there's also something called mean squared error. $E((\hat{\theta}-\theta)^2)$
- Sometimes unbiased estimators have *higher* mean squared error than biased ones.
- MSE $= Var(\hat{\theta}) + [Bias(\theta, \hat{\theta})]^2$
- So what makes a "good" estimator is sometimes in question.
- Indeed, uncorrected (biased) sample variance has lower mean squared error! (google for a proof)
- BLUE: best, linear, unbiased estimator.
- Will see the above a lot when folks talk about regression.

9 Conditional Association, Regression

- Simpson's Paradox is when an association in the data disappears when groups are combined.
- In effect, this is about an association disappearing when something else is, in effect, conditioned on.
- In the Berkeley case, sex bias in the data overall was not accounting for which disciplines men and women disproportionately applied to. Programs like the English department, which in the data attracted a disproportionate number of female applicants, admitted very few people as a percentage.
- We can also talk about this in the context of regression models. *Do you have a sense of what regression is, intuitively? I think this was talked about a bit in math camp? Essentially, drawing a line through stuff. Or drawing a plane. Or draw a hyperplane. Who wants to draw a hyperplane?*
- Intuition from the basics: what are we doing when we look at a single-variable regression, i.e. a model of form $y = \beta_0 + \beta_1 x_1 + e$?
- Question: So if all we are interested in is the effect of x_1 on y, why don't we just do this all the time?
- Model: Drowning deaths $= \beta_0 + \beta_1$ice-cream sales $+ e$ What's the issue?
- "Lurking" variables, or omitted variable bias. Classic case: where a dependent variable of interest is related to some other dependent variable *and* the independent variable.
- Quick side note: what does it mean for an estimator to be biased?
- Question: Suppose the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$. Say we initially estimate $y = \beta_0 + \beta_1 x_1$. Will including x_2 reduce the bias in our estimate of β_1?
- Follow up question: what are the conditions under which the above has different answers?
- Say A causes B causes C. Should we include both A and B?
- In any event, we need a mechanism of "controlling" for omitted variables that we feel may be confounding our analysis.
- Multiple regression does this, in some sense. We "partial out" the effects of other independent variables in order to isolate the effect of a single variable.

- Note: we can obtain estimates for β_1 in a two variable model in a way that illustrates the partialling out interpretation well. Regress x_1 on x_2, obtain the residuals r_1, then regress y on thse residuals r_1. This will give the estimate of β_1 when the effects of x_2 have been partialled out, and is the same as what we would have gotten from doing multiple regression in the first place (though not the same standard errors). See code at end of notes to try it out.
- Predicting bias direction with two variables:

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

- Things get more complicated when the true model includes more than two variables. For instance, say you have x_3 which is uncorrelated with x_1 but is correlated with x_2. Does not including x_3 induce bias in our coefficient for x_1?
- Answer is yes, if x_2 is correlated with x_1. Doesn't matter that x_3 is not directly correlated with x_1.
- Say our correct model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Say we start off with just x_1, and say that we know both x_2 and x_3 are correlated with both x_1 and y. Does our bias decrease when we go from including only including x_1 to including x_1 and x_2?
- The answer: not necessarily! This is the subject of Kevin Clarke's wonderfully-titled paper: *The Phantom Menace: Omitted Variable Bias in Econometric Research.*
- If x_2 introduces negative bias and x_3 introduces positive bias, then including only one and not the other means you could be further from the truth than with neither.
- As a result, unless we have the fully specified model, we can't even know if including a variable that belongs in the model with increase or decrease the bias on the coefficient estimate of interest.
- Summary question: You are interested in the effect of x_1 on y. x_2 is also part of the true model. Should you include it?

9.1 Practice with matrix algebra

- What is $(\boldsymbol{y} - \boldsymbol{X\beta})'$?
- Important matrix transpose properties: $\boldsymbol{(A')' = A}$
- Additive: $\boldsymbol{(A + B)' = A' + B'}$
- $\boldsymbol{(AB)' = B'A'}$
- $\boldsymbol{(cA)' = cA'}$
- $\boldsymbol{(A^{-1})' = (A')^{-1}}$
- What is $\boldsymbol{ee'}$ versus $\boldsymbol{e'e}$? (Think, $n \times 1$ and $1 \times n$ versus $1 \times n$ and $n \times 1$).
- $\boldsymbol{e'e}$ is sum of squared residuals! This is, indeed, what we minimize for OLS.
- Standard error of regression: $\hat{\sigma} = \sqrt{\frac{\boldsymbol{e'e}}{n-k}}$. This is, essentially, a measure of how precise the predicted values from the regression are. Can be used to compute a range around our predicted values which we would expect the actual values to fall within with some probability (e.g. 95% confidence). Other interpretation: variance of the residuals.
- More often, we're interested in the standard errors of the coefficient estimates. This is what allows us to make statements like "the values of this coefficient is statistically different than zero", or, as it is more often stated, "the effect of x on y is statistically significant".

- To get this variance covariance matrix, find $\hat{\sigma}^2(\boldsymbol{X'X})^{-1}$.
- This will give covariances on the off-diagonal, with variances for each coefficient estimate $\hat{\beta}_0, \hat{\beta}_1, \ldots$ along the diagonal.
- Take the square roots of these to get the standard errors of the coefficient estimates.
- Recall from math camp we did this:
  A matrix is idempotent if multiplying it by itself returns the same matrix (i.e. $\boldsymbol{AA} = \boldsymbol{A}$). Prove that $\boldsymbol{I - X(X'X)^{-1}X'}$ is idempotent.
  Ans:
  $\boldsymbol{(I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X')}$
  $= \boldsymbol{II - 2IX(X'X)^{-1}X' + X(X'X)^{-1}X'X(X'X)^{-1}X'}$
  $= \boldsymbol{I - 2X(X'X)^{-1}X' + X(X'X)^{-1}X'}$
  $= \boldsymbol{I - X(X'X)^{-1}X'}$
- The above matrix is called the "residual maker", often denoted $\boldsymbol{M}$. Why might this be the case?
- Recall that the regression equation gives us a best fit for $\boldsymbol{y = X\beta}$ by choosing $\boldsymbol{\beta}$ to minimize the sum of squares.
- The OLS estimate of $\boldsymbol{\beta}$ is $\boldsymbol{(X'X)^{-1}X'y}$, as you will show in question 1 of the problem set. So what happens if you compute $\boldsymbol{My}$?
- Answer: You get the residuals!

9.2 OLS estimator derivation

Ordinary least squares linear regression is based on minimizing the squared differences between your regression "line" (hyperplane) and your observed data. Same deal as what we did earlier with least squares estimators for the mean. So, want to minimize $e'e$ where $e = y - XB$ (can you see why this equation holds?)

$$\begin{aligned} min_B\ (y - XB)'(y - XB) &= (y' - B'X')(y - XB) \\ &= y'y - B'X'y - y'XB + B'X'XB \\ &= y'y - 2B'X'y + B'X'XB \end{aligned}$$

taking derivative with respect to B and setting to zero returns

$$\begin{aligned} -2X'y + 2X'XB &= 0 \\ \leftrightarrow X'XB &= X'y \text{ (Note, this is sometimes called the normal equation(s))} \\ \leftrightarrow (X'X)^{-1}X'XB &= (X'X)^{-1}X'y \\ \leftrightarrow B &= (X'X)^{-1}X'y \end{aligned}$$

9.3 Omitted variable bias and partitioned regression

Let's see how to mathematically represent omitted variable bias using partitioned matrices. First, let's consider the initial "normal equation" for regression, but with a data matrix I will label $\boldsymbol{X_1}$ for reasons that will become obvious after.

$$\boldsymbol{X_1'X_1b_1 = X_1'y}$$

Which leads to the regression equation to solve for b:

$$\boldsymbol{b_1 = (X_1'X_1)^{-1}X_1'y}$$

Now, let's say that $\boldsymbol{X_1}$ includes all the variables except one, denoted $\boldsymbol{X_2}$. Now say we want to add this in. We can represent this in a partioned matrix like so.

$$\begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{X_2}\end{bmatrix}' \begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{X_2}\end{bmatrix} \begin{bmatrix}\boldsymbol{b_1} \\ \boldsymbol{b_2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{X_2}\end{bmatrix}' \boldsymbol{y}$$

Components of partitioned matrices can mostly be treated the same way as you would treat elements of regular matrices, while keeping in mind a few things when you do things like transposing or finding inverses.

$$\begin{aligned} &\leftrightarrow \begin{bmatrix}\boldsymbol{X_1'} \\ \boldsymbol{X_2'}\end{bmatrix} \begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{X_2}\end{bmatrix} \begin{bmatrix}\boldsymbol{b_1} \\ \boldsymbol{b_2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X_1'} \\ \boldsymbol{X_2'}\end{bmatrix} \boldsymbol{y} \\ &\leftrightarrow \begin{bmatrix}\boldsymbol{X_1'X_1} & \boldsymbol{X_1'X_2} \\ \boldsymbol{X_2'X_1} & \boldsymbol{X_2'X_2}\end{bmatrix} \begin{bmatrix}\boldsymbol{b_1} \\ \boldsymbol{b_2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X_1'y} \\ \boldsymbol{X_2'y}\end{bmatrix} \end{aligned}$$

Let's start to multiply out to try to solve for $\boldsymbol{b_1}$, solving for the part of the left hand side equal to $\boldsymbol{X_1'y}$

$$\begin{aligned} \boldsymbol{X_1'X_1b_1} + \boldsymbol{X_1'X_2b_2} &= \boldsymbol{X_1'y} \\ \leftrightarrow \boldsymbol{X_1'X_1b_1} &= \boldsymbol{X_1'y} - \boldsymbol{X_1'X_2b_2} \\ \leftrightarrow (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_1b_1} &= (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'y} - (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2b_2} \\ \leftrightarrow \boldsymbol{b_1} &= (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'y} - (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2b_2} \end{aligned}$$

What do you notice about $(\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'y}$? It's the same as the initial regression estimator when we didn't have the omitted variable! So when will our estimate for $\boldsymbol{b_1}$ be the same with the added variable as it was without it? When $(\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2b_2}$ is equal to zero. Let's unpack the components of this. $(\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2}$ is a regression of $\boldsymbol{X_2}$ on $\boldsymbol{X_1}$, and $\boldsymbol{b_2}$ is a measure of the effect of $\boldsymbol{X_2}$ on $\boldsymbol{y}$. This is all very similar to our initial bias table. Note that $\boldsymbol{X_2}$ can be generalized to a set of omitted variables instead of just one omitted variable.

9.4 Alternative approach to omitted variable bias

- Consider if you have some set of regressors contained in $\boldsymbol{X}$ and then an additional variable contained in $\boldsymbol{z}$ which is part of the true model but isn't included in your regression.
- The true linear model looks like $y = \boldsymbol{X\beta} + \boldsymbol{z\alpha} + \epsilon$.
- Consider that the regression estimator is $\hat{\boldsymbol{\beta}} = \boldsymbol{(X'X)^{-1}X'Y}$. This is unbiased are when $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
- So let's consider $E(\boldsymbol{(X'X)^{-1}X'Y})$. Plug in for y to get:
  $E[\boldsymbol{(X'X)^{-1}X'}(\boldsymbol{X\beta} + \boldsymbol{z\alpha} + \epsilon)]$.
- So think about expanding that out, and separating the different terms (by linearity of expectation). Then consider what each part is equal to. Under what conditions will the full expression equal β? Ultimately, this will end up looking similar to the approach from using partitioned matrices.

9.5 Quick additional note on partioned matrices

- Say you want to show that one variable in a data matrix was a scalar multiple of the other and show the implications.
- It is straightforward to show the result we want (that making one of the variables a scalar multiple means we can't invert $\boldsymbol{X'X}$) if we use a simple partioned matrix.
- $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x_1} & r\boldsymbol{x_2}\end{bmatrix}$, $\boldsymbol{X'} = \begin{bmatrix}\boldsymbol{x_1'} \\ \boldsymbol{x_2'}\end{bmatrix}$

- If $\boldsymbol{x_2} = r\boldsymbol{x_1}$ then $\boldsymbol{X'X} = \begin{bmatrix} \boldsymbol{x_1'} \\ r\boldsymbol{x_1'} \end{bmatrix} \begin{bmatrix} \boldsymbol{x_1} & r\boldsymbol{x_1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x_1'x_1} & \boldsymbol{x_1'}r\boldsymbol{x_1} \\ r\boldsymbol{x_1'x_1} & r\boldsymbol{x_1'}r\boldsymbol{x_1} \end{bmatrix}$
- Which means that $det(\boldsymbol{X'X}) = \boldsymbol{x_1'x_1}r^2\boldsymbol{x_1'x_1} - r^2\boldsymbol{x_1'x_1x_1'x_1} = 0$, and thus it's not invertible.

10 Sampling Theory

- Before we get to discussing hypothesis testing and confidence intervals, it's worth taking a moment to discuss a few results from sampling theory that end up being implicitly evoked when we do stuff like finding t-statistics or z-statistics and comparing them to z tables and whatnot. It's also good to know the jargon.
- When we say a "random sample", this implies that we are talking about a sample where each observation is independent and identically distributed. We will focus on random samples.
- Sampling theory is, generally speaking, about properties of the *sample mean*. It is important to keep this in mind when you're doing testing.
- Sample means are just the average of the sample. You probably know this already.
- The big theorems from sampling theory are the law of large numbers (LLN) and the central limit theorem (CLT).
- There are, in fact, several "laws of large numbers" (LLNs), though people often refer to "the" law of large numbers.
- The main differences between them are in the assumptions they make to get certain results. This will usually not be too important to you in practical applications.
- Kolmogorov's Strong LLN states that if you have random samples of size n, if the mean of the distribution exists (call it μ), then $\bar{X}_n$ will converge to μ as n goes to infinity.
- Stated differently: $plim(\bar{X}_n) = \mu$
- Essentially, this means that as your sample size gets very large, your sample mean is going to get super close to the actual mean. As the sample size went to infinity, you'd end up at the actual mean.
- This is a good start for thinking about why we might want more data; getting more seems to get use closer to the correct answer (or at least, gets us to the right answer if n goes to infinity, which is technically different, but probably not in ways you should focus too much on).
- However, practically this doesn't give us a way of quantifying our uncertainty about how close we are to the correct answer. For this, we might want to know more about the distribution of the sample mean.
- For this we have Central Limit Theorems (CLT).
- We'll just focus on the Lindeberg-Levy CLT, which has as assumptions that your sample is iid, and you have finite mean and variance.
- These theorems aren't stated in terms of the sample mean directly, although they can be indirectly interpreted as having implications for the sample mean.
- Instead, what they state is that $Z_n = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$
- Or in words, the scaled and standardized version of the sample mean converges to a normal distribution with mean zero and variance 1 (i.e. a standard normal).

- Now in some sense, if you didn't do the standardizing and scaling, you'd expect the unscaled $\bar{X}_n$ to converge to $N(\mu, \sigma^2/n)$, right? So why are we bothering to state things in terms of the standardized one?
- The above is mostly right, but keep in mind that all these properties are stated in terms of what happens as n goes to infinity. As we noted in the LLN, as n goes to infinity the sample mean converges to a constant, which does not in fact have any distribution!
- To see this, compare the two rows of this simulation I did, between taking sample means as the sample size increased (first row), to taking standardized and scaled sample means as the sample size increased (second row).

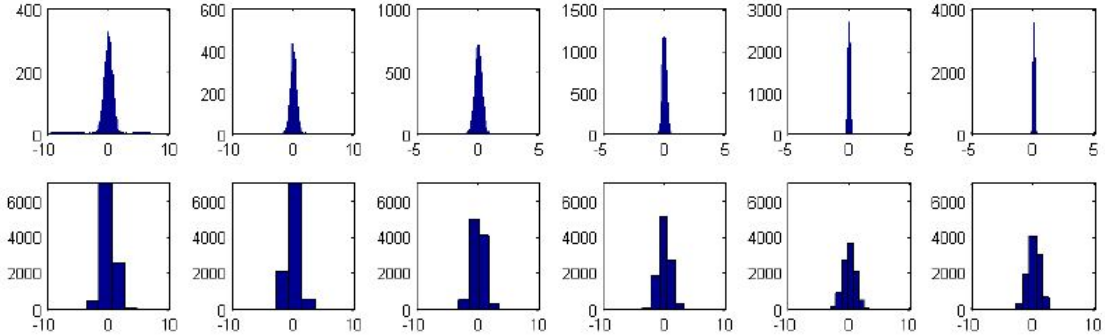


Figure 1: Sample means versus z statistics

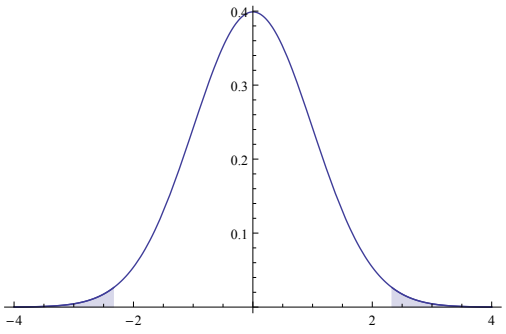
- You can see that the sample means collapse to a point, whereas the z statistics (i.e. the scaled and standardized sample means) start to look more normal.
- So to state the CLT in terms of asymptotics (i.e. properties as n goes to infinity) we need to do the scaling, as technically, the distribution of the sample mean is degenerate, i.e. $\bar{X}_n \xrightarrow{d} \mu$.
- It has the added advantage of this: whenever we compute a z-statistic, we are adjusting it in such a way so that it can be compared to a *single* distribution from the normal family of distributions, i.e. the standard normal distribution.
- This is what you're doing when you check a z-table for probabilities. The z-table is just giving you different probabilities from a standard normal distribution.
- The Central Limit Theorem is what allows you to do this at high n, regardless of what the original distribution looked like. Even if it was super weird looking (i.e. not at all normal looking), the CLT tells you that the z-statistic (scaled and standardized sample mean) will look like a standard normal distribution as n goes to infinity.
- This is super powerful! Most of the stuff you do in standard statistics classes is based on this.
- The other thing you'll hear about a lot is the t-statistic. People will tell you this is the thing to use when you don't have the true variance, and when your sample size is small.
- Looks like $t_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s}$. So like a z-statistic, but with the true variance replaced with the sample variance.
- If you have the true variance, jump right in to using the z-statistic.

- If you have only the sample variance, then the t statistic is really only helpful if your source distribution is normal. If the source distribution is normal, then your t-statistic will have a t-distribution with $n-1$ degrees of freedom.
- However, if the source distribution is non-normal, then there's no reason that your t-statistic will be t-distributed. So this is pretty restrictive.
- The good news is this. As n goes to infinity, the t-distribution converges to the standard normal distribution. Moreover, the CLT does tell us that as n goes to infinity the sample variance will be a fine substitute for the true variance, so we can rely on normal theory results.
- Basically, the t statistic only gives us anything different from the z-statistic for small sample sizes (say $n < 30$), and where the source distribution is normal. If the source distribution is nonnormal and we have $n < 30$, then the CLT probably can't help us, but neither can the t-statistic. If n is very high, then the t-statistic doesn't help us, because the CLT results about the z statistic get us all the way there without worrying about the difference between sample variance and true variance.
- However, people will generally use the t statistic in most cases. This is because, while it rarely helps much over the z statistic, it's identical for high n, and only differs at low n, where either the source distribution is normal, so the t-statistic helps, or the source distribution is nonnormal, and you're screwed anyway.
- Basically, the t statistic strictly dominates (to use game theory jargon for any of you who know it).

11 Hypothesis Testing

- The Null Hypothesis! This depends on what is the alternative claim to your actual hypothesis.
- For instance, if our hypothesis is that X has an effect on Y, then the null hypothesis, intuitively, would probably be that X has no effect on Y.
- Once we have an estimate, the p-value is generally interpreted as the probability we would obtain a particular test statistic (estimate) at least as extreme given that the null hypothesis is true.
- One-tailed and two-tailed hypotheses. E.g. x greater than zero versus x different than zero.
- Computing p-values for two-tailed hypotheses entails "doubling" the probability in one-tailed.
- P-values can be used to determine whether you reject the null. However, it is problematic to think of p-values as determining the confidence level at which you should conduct your test.
- Type I error is when you reject the null hypothesis when it is true.

- Type II error is when you do not reject the null hypothesis when it is false.
- Ye old analogy: Type I: Convicting an innocent person. Type II: Not convicting a guilty person.
- Intuitively, everything else held constant, reducing the probability of Type I error is inevitably going to increase the probability of Type II error.
- If our null hypothesis is something like $x = 0$, we need to consider what the sampling distribution of a statistic used to estimate x would look like.
- This is where all the sampling theory stuff comes in. Think back to the MLE problem set: a lot of the parameter estimators were just sample means.
- The sampling theory stuff can often be extended to talk about distributions of more complicated estimators. For the next problem set, you're only dealing with sample means.
- Once we establish a reference distribution for our test statistic under the null hypothesis, we determine a "rejection region", which is a set of values where we reject the null hypothesis. This is our decision rule.
- If we suitably scale and standardize our test statistic, our reference distribution for the test statistic can usually be something like the standard normal or the the t distribution.
- Keep in mind the sampling theory stuff. For instance, this stuff tells us that if $\mu = 0$, then $Z_n = \frac{\sqrt{n}(\bar{X}_n - 0)}{\sigma}$ will be approximately standard normal distributed at high n. We can compare whatever test statistic we find (using $\mu = 0$) to this distribution we get from hypothesizing that $\mu = 0$.
- Similarly, if our null was that $\mu = 7$, we would expect $Z_n = \frac{\sqrt{n}(\bar{X}_n - 7)}{\sigma} \xrightarrow{d} N(0, 1)$
- The probability that we have Type I error is computed by finding the probability that our sample statistic would fall in the rejection region, given that the null hypothesis is true.
- Oftentimes, we set the probability of Type I error first and then determine the rejection region to fit this.
- Let's do an example. Say our null hypothesis is that $\mu = 3$. Let's say we want to do this two-tailed test at 98% confidence. So we want a z-statistic where the probability of Type I error is $\alpha = 0.02$, which implies a probability of $\alpha/2 = 0.01$ at each tail. With a standard normal distribution, this implies a rejection region of $z < -2.33$ and $z > 2.33$, as below.

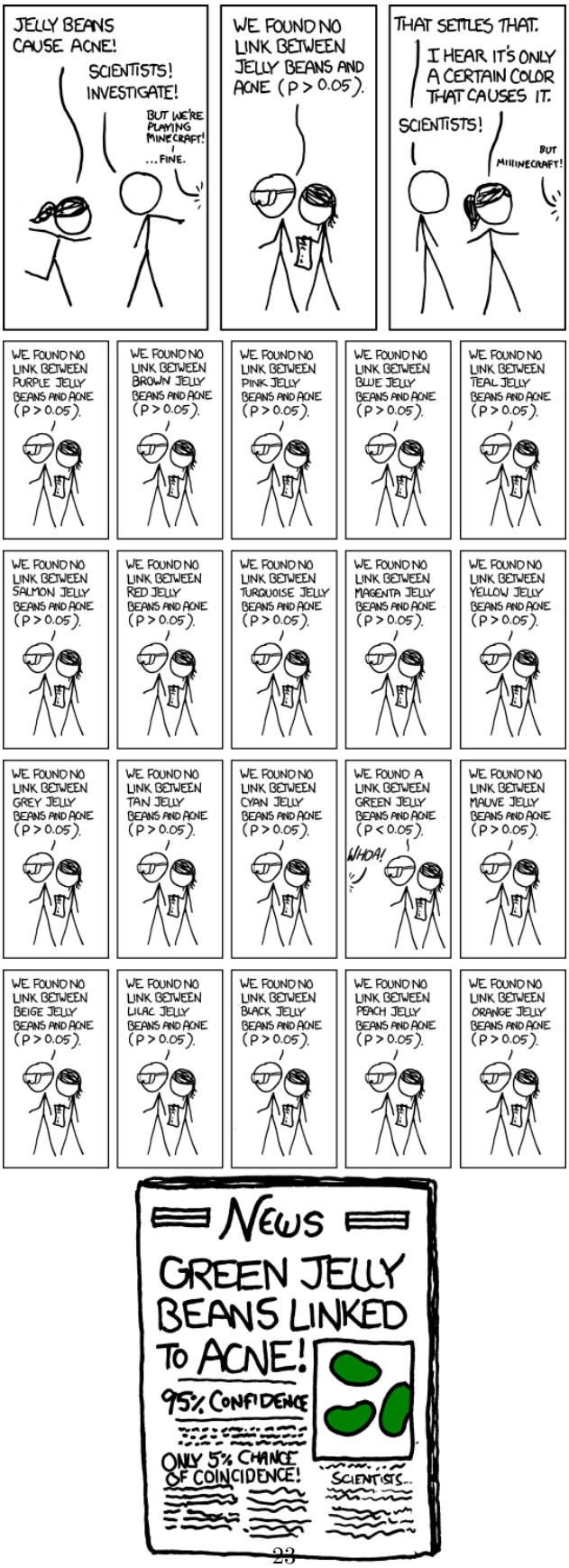


- Now say we have a sample of size $n = 100$, and we get $\bar{X}_n = 1$ and $s^2 = 25$. This is probably sufficiently high n to assume the z statistic would be approximately normally distributed. So we compute it, under the assumption that the null is true: $Z_n = \frac{\sqrt{100}(1-3)}{5} = \frac{-20}{5} = -4$.

- This z-statistic is well within our rejection region, and gives us a p-value of something like 0.00006. So we would emphatically reject the null hypothesis that $\mu = -3$.
- But let's say our null hypothesis was, instead, that $\mu = 0$. This is, of course, a very common null. Our statistic is now $\frac{\sqrt{100}(1-0)}{5} = 2$. We now have a p-value of 0.0455 and would not reject the null hypothesis at 98% confidence, though we might expect to reject it at lower confidence levels.
- In fact, we know that we would (barely) reject the null at 95% confidence ($\alpha = 0.05$) given that the p-value is less than 0.05. However, it would be a mistake to post-hoc determine the testing level based on our p-value. This is something that should be determined ex ante.

11.1 Multiple Testing

- Something Walter discussed in class is multiple testing.
- Consider that if you do 20 tests at 95% confidence, you would get one "false positive" on average from those tests, insofar as we expect to see statistics so extreme as to be in the rejection region when the null is true only one time out of 20.
- More precisely, if we are testing at 95% confidence ($\alpha = 0.05$), the probability of *not* getting a false positive (i.e. not rejecting the null when the null is true) is 19/20. So the probability of not getting any false positive in any test of 20 is $(19/20)^{20} = 0.3584$. This implies that there is a $1 - 0.3584 = 0.6416$ chance of getting at least one false positive in those 20 tests.
- If we are concerned about controlling Type I error (i.e. the error of false positive) over multiple tests, we may need to perform a correction.
- Familywise Error Rate (FWER) is the probability of rejecting at least one true null hypothesis over a series of tests, i.e. the probability of making at least one Type I error, i.e. the probability of one false positive.
- To use the old metaphor: this would be analogous to controlling the probability of sending *any* innocent people to prison, over, say, 20 criminal cases. If we simply adopted an evidentiary standard that, in any one case, gives a 5% chance of sending an innocent person to prison, we would end up with a lot of innocent people in prison when you have multiple cases.
- If we want to control the probability of a false positive to be $\leq \alpha$, one way to do that is to divide the testing level for each test by the number of tests. So if we were doing 20 tests, and wanted to control the probability of getting at least one Type I error to be less than α, we would do each test at $\alpha/20$.
- Consider 20 tests, where we want the probability of getting at least one Type I error to be less than $\alpha = 0.05$. We would do tests at $\alpha/n = 0.05/20 = 0.0025$
- This implies the probability for any one test of *not* committing a Type I error is $1 - 0.0025 = 0.9975$.
- As before, we take $(0.9975)^{20} = 0.95117$. If we subtract this from 1, we get $1 - 0.95117 = 0.04883$. So approximately 0.05.
- So this is a simple correction in which dividing α by 20 allowed us to control the probability of getting one or more false positives to be less than α.
- This is called a Bonferroni correction, and is one way of dealing with false positives in multiple testing.
- There is an interesting question about when to use these corrections. Over one's entire career? Throughout a discipline? In practice, they are probably rarely used. You might consider using them if you have a joint hypothesis which implies that multiple things must be true simultaneously. But even then, there are often issues, and sometimes better ways of going about it.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

12 Confidence Intervals

- In principle, this should give us an interval around our estimate which has a high probability of containing the unknown parameter.
- What's the probability that the parameter is equal to our parameter estimate? Zero, if the parameter space is continuous!
- Confidence intervals essentially tell us that if we were to generate a series of confidence intervals of particular sizes, some percentage of these would contain the parameter.
- The percentage of intervals that would contain the parameter (on average) is based on what level of confidence our interval is.
- The level of confidence (e.g. 95% confidence interval) refers to the procedure by which this interval is constructed.
- 95% confidence intervals: what fraction of such intervals will contain the parameter? How about 90% confidence? 99%?
- The parameter does not have x% of falling within the interval. It either does or does not, insofar as we are in the world where parameters are constants, and thus do not themselves exhibit randomness.
- We need to have knowledge of the distribution of our sample statistic in order to construct these intervals.
- Normal theory suggests that for sufficiently high n, we can sometimes assume that the sampling distribution of our estimator is approximately normal.
- When we don't have the variance, and have to use an estimate of variance, we may want to use t-statistic/distribution, depending on how high n is, and whether or not we buy that the original data was normal.

12.1 Computing Confidence Intervals For Our Parameter Estimates

We have found some statistic by which we estimate the parameter value of the distribution. We now need to consider the sampling distribution of that statistic, so that we can construct confidence intervals. Note that a parameter, in the standard interpretation, does not have a distribution, as it is fixed, but the statistic that estimates that parameter does. Example:

$$\begin{aligned}\hat{\mu} &= Var(\frac{\Sigma x_i}{n}) \\ &= \frac{1}{n^2} Var(\Sigma x_i)\end{aligned}$$

Given that each x_i is i.i.d

$$\begin{aligned}&= \frac{1}{n^2} \Sigma Var(x_i) \\ &= \frac{1}{n^2} \Sigma \sigma^2 \\ &= \frac{n\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n}\end{aligned}$$

So let's do a quick example, using similar numbers to what we had in the hypothesis testing section. I.e. let's say we have a sample of size $n = 100$, and we get $\bar{X}_n = 1$ and $s^2 = 25$. So $\bar{X}_n = 1$ is our estimate of the mean. Now let's say we want to construct a 95% confidence interval around this estimate. If the sampling statistic is likely to be normally distributed around the true mean, then for any given estimate the parameter is just as likely to be higher than the estimate as it is lower. We construct the interval:

$$\bar{X}_n \pm Z_{0.05/2} \cdot \sqrt{\frac{25}{100}} = 1 \pm 1.96 \cdot 0.5 = [0.02, 1.98]$$

There is a theoretical way to connect this to hypothesis testing. Note that the 95% confidence interval barely does not overlap zero. This suggests that, at 95% confidence, one would (barely) reject a null hypothesis that $\mu = 0$. This corresponds with what we found with the same numbers in the hypothesis testing section.

13 GLMs/Statistical Modeling Generally

- Wrapping link distribution/function around linear model
- E.g. with binary models, where you want probabilities bounded between zero and one.
- Logit and probit. Logit uses logistic function, probit uses normal distribution. Both often have similar properties.
- In traditional linear model, if say, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, then $\frac{\partial y}{\partial x_2} = \beta_2$, and thus the coefficients have a natural interpretation as the "marginal effect" or "marginal association" between two variables.
- How to report marginal effects when you have a function wrapped around the linear model?
- If you have $y = F(\boldsymbol{X}\boldsymbol{\beta})$, then $\frac{\partial y}{\partial X} = F'(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\beta}$
- Logit takes form of $p_i(1-p_i)\beta_j$. Still doesn't give us one single marginal effect, as this will vary across the values of the covariates.
- As a general matter, this makes the coefficients more difficult to interpret.
- Techniques that are used when reporting marginal effects in papers include setting all other variables to their means, or as an average of marginal effects at each observation.
- R package "erer" can be used to compute these, with function maBina.
- The stuff generally referred to as GLMs are just examples of statistical models with particularly specified structural forms.
- These models are not actually linear, for what it's worth.
- Other examples. Truncated/censored models: Tobit. Need a distribution with a mass point.
- Duration models: how long will someone be likely to stay in power? Or, far more interestingly, how long before a lightbulb burns out?
- All these involve making strong distributional assumptions
- Third class in sequence used to be called MLE, because most of these would be estimated via MLE. Maybe MM (or GMM) sometimes. So advanced techniques were all about learning new models for dealing with different kinds of data.
- These days, many folks are moving towards techniques that don't make as many distributional assumptions, or are eschewing the modeling enterprise entirely, looking more towards quasi-exogenous shocks and whatnot.

14 Contingency Tables

- Contingency tables are just matrices that contain the frequencies of the different variables.
- Can be used to look for associations between variables. Consider the following:

	Canadian	American	Both
Jerk	5	200	205
Not Jerk	45	200	245
Total	50	400	450

- We might be curious about whether jerkiness is associated with being American or not. To assess this, we can assess the observed proportions against what we would expect if there was no relation.
- If there was no relationship, we just take whatever the population proportion is off jerks (e.g. $\frac{205}{440} = 45.5\%$ Jerks), and multiply by the total number in each category (50 canadians, 400 americans). This gives us our "expected" frequencies in absence of a relationship. We get:

	Canadian	American	Total
Jerk	22.7	182	205
Not Jerk	27.3	218	245
Total	50	400	450

- Just from eyeballing it, we can see this is pretty different from the actual observed frequency distribution. However, we can do a more formal test: the chi-squared test.
- The chi-squared statistic is: $\sum \frac{(Actual-Expected)^2}{Expected}$.
- So in the case, we have: $\frac{(22.7-5)^2}{22.7} + \frac{(27.3-45)^2}{27.3} + \frac{(218-200)^2}{218} + \frac{(182-200)^2}{182} = 28.54$
- Compare this against a chi-squared distribution, with $(r-1)(c-1)$ degrees of freedom, where r is number of rows, and c is number of columns (not including the "Total" row and column).
- So this would have $(2-1)(2-1) = 1$ degree of freedom.
- Gives us a p-value of approximately zero. So clearly there is a relationship!
- To compare against some arbitray proportion, just plug that in for the expected frequencies. Similar to what you do with other hypothesis testing.
- If you want to consider whether a relationship holds up when conditioned on some third variable, you could consider doing two separate tests with the data split by that variable.
- Standard assumptions of random sampling from the distributions for the test to work.
- In terms of assessing causality: what do you think?
- You could talk about what the conditional effect would look like under different values of the variable by comparing proportions.