BIOMED SCI 552:

STATISTICAL THINKING

LECTURE 6: ESTIMATION

QUESTIONS FROM TUESDAY?

WHAT IS ESTIMATION?

- We've alluded to estimation a lot in this class so far, because all these concepts are somewhat intertwined
- The core of much of what we want to *do* with statistics is called estimation
- Estimation is the process of inferring an unknown (and unknowable) quantity from the population using data from the same
- It is, in effect, our best guess using data

SO WHAT'S AN ESTIMATOR?

- An estimator is the method by which we obtain an estimate
- These can be quite sophisticated, or quite simple
- One of the ones we've already used in this class a lot is the estimator of the sample mean, $\bar{X}$, which is an estimate of the population mean
- $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

CONSISTENT ESTIMATORS

- A consistent estimator (sometimes called an asymptotically consistent estimator) is one where as sample size increases, you will eventually converge on the true estimate
- An example of an inconsistent estimator would be an estimate of the population mean that has the first value you sample
  - Technically, this is an unbiased estimator
  - But it does not get closer to the population estimate because it is fixed regardless of sample size
- Encountering an inconsistent estimator in the wild is *rare*

EFFICIENT ESTIMATORS

- This just means that there's no *other* estimator that would arrive at the same estimate with a smaller accompanying variance

HATS AND BARS

- $\hat{X}$ and $\bar{X}$ are both going to appear today
- They feel like they often mean the same thing, but they don't
- $\bar{X}$ is the sample mean of something
- $\hat{X}$ is an estimate
- We can use $\hat{X}$ as an estimate of $\bar{X}$, but it's not the only one

BIASED VS UNBIASED ESTIMATORS

- The bias of an estimator is the difference between the estimator's expected value and the true value of the parameter being estimated
- For an unbiased estimator, this difference is 0
- For a biased estimator, the difference is either <0 or >0

- Why would we ever use a biased estimator?

SOMETIMES AN UNBIASED ESTIMATOR DOESN'T EXIST

- The is, for example, no unbiased estimator for $\frac{1}{\mu}$ with observations from a Poisson distribution with a mean of μ

BIAS-VARIANCE TRADE OFF

- It is possible, and indeed quite common, that an *unbiased* estimator may also be considerably less precise
- If we think about the true population value we're estimating as the target, we can measure how "off" we are by looking at "Mean Square Error"
- $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$
- It can be shown that this is equal to:
- $MSE(\hat{\theta}) = (Bias(\hat{\theta}, \theta)^2 + Variance(\hat{\theta})$

BIAS-VARIANCE TRADEOFF

- This means that there are often situations where it might be desirable to trade a little bit of bias for a large reduction in variance, and still end up close to your answer
- While we said that high bias, low variance is a bad place to potentially be, smallish bias, low variance may be preferable to unbiased and high variance

AN EXAMPLE…

- Lets say I tell you I'm thinking of a number, and the closer you get to the number, the greater your prize
- I tell you that to pick your number, you have to choose between one of two methods to randomly draw a guess (i.e. I let you choose between two estimators):
  - $\hat{X} \sim Uniform(X - 5, X)$
  - $\hat{X} \sim Uniform(X - 100, X + 100)$
- Which would you prefer?

IN A PERFECT WORLD

- Obviously, in a perfect world, we would prefer an unbiased estimate with low variance, but we cannot always promise this will be true
- Sometimes, known bias in an estimator can be corrected, etc.
- Many, if not most, estimators are biased
- A deeper dive into this gets very complicated very quickly
  - For the most part, the methods we use are at the very least ones with known properties

HOW THIS COMES UP IN THE BIOMEDICAL SCIENCES

- There are two major ways this comes up in the biomedical sciences
- Pure estimation
  - A colleague of mine: "Epidemiologists just…*measure things*."
  - Very commonly estimation is part of a larger chain of research
  - Examples:
    - Estimating the average patient length of stay to model hospital bed availability
    - Estimating the prevalence of a particular condition for allocating resources for it
    - Estimating a bacterial growth rate under certain conditions
  - Here, the goal is to minimize MSE (or some other measure of overall error), *not* just to minimize bias

”CONTROLLING FOR VARIABLES”

- When we adjust for things, why don’t we just adjust for everything we can possibly think of?

"CONTROLLING FOR VARIABLES"

- When we adjust for things, why don't we just adjust for everything we can possibly think of?

- The more variables we add, the more variance increases
- We are essentially "spending" variance to protect us from bias
- At some point there's a diminishing payoff for this
  - There are formal ways to assess this

THE BAYESIAN PERSPECTIVE

- Gelman et al., *Bayesian Data Analysis*
- "From a Bayesian perspective, the principle of unbiasedness is reasonable in the limit of large samples but otherwise is potentially misleading. The major difficulties arise when there are many parameters to be estimated and our knowledge or partial knowledge of some of these parameters is clearly relevant to the estimation of others. Requiring unbiased estimates will often lead to relevant information being ignored... In sampling theory terms, minimizing bias will often lead to counterproductive increases in variance."

IS A PRIOR BIAS?

CONNECTING THIS TO SAMPLING

- All of this falls under what's called "sampling theory" because it's centered on what inference we can draw from a sample
- Our estimator itself cannot be ensured to be unbiased
- There is often a notion that "Why can't we just make a method that doesn't have *X* undesirably property?"
- Sadly, math doesn't work like that

CONNECTING THIS TO SAMPLING

- All of this falls under what's called "sampling theory" because it's centered on what inference we can draw from a sample
- Our estimator itself cannot be ensured to be unbiased
- There is often a notion that "Why can't we just make a method that doesn't have *X* undesirably property?"
- Sadly, math doesn't work like that

CONNECTING THIS TO SAMPLING

- Given we cannot realistically control the properties of the estimators we use…
- We should endeavor to minimize the bias that comes from things we *can* control
- Our goal is to ensure that we do not add any more bias into a sample than what is inevitable due to problems with sampling frames, estimators, etc.