Rayanna Harduarsingh

September 16th, 2020

Module 4 Notes

IST 687

Chapter 10/Lecture 4.1: Sample in a Jar

- **Sampling distributions are the conceptual key to statistical inference.**
    - Example: Imagine a jar full of 100 red and 100 blue gumballs. But, when they were all poured in, they all got mixed in together. If you drew 8 random gumballs, what colors would you get? It would not work out evenly (4 red and 4 blue), or maybe in a rare instance. This is an example of a sampling distribution and learning from this gumball sample can help us **understand patterns or predict outcomes in the long run.**
- **Sampling in R:**
    - Sample() function
    - "sample(USstatePops$april10census, size=8, replace=TRUE)"
- **Question: Why is sampling from a population important? What are some key things to think about when sampling?**
    - Sampling from a population is important because it cuts the time of research almost in half. Researching an entire population is not impossible, but very complex and difficult. It's a massive approach to do when conducting a project or research. Therefore to make the research process easier, it is important sample a population. A sample is a small piece that represents something larger. When sampling, it's important to not be bias. It's important to keep the sampling as random as possible to provide more accurate results and avoid bias. However, there are times when your sample can't be random, and your objects need to have a common core. Also, when picking a sample, it's important research before to know what/who your sampling to better serve your objective.
- **Replicating Samples**

- Not interested in any one sample, but in what happens over the long run. We have R to repeat this process for us.
    - Replicating Samples in R:
    - replicate() function:
        - replicate(4, mean(sample(USstatePops$april10census, size=8,replace=TRUE)),simplify=TRUE)
        - mean(replicate(400,mean(sample(USstatePops$april10census, size=16, replace=TRUE)), simplify=TRUE))
            - Interpretation:
                - Draw 400 samples of size 16 from our state population.
                - Calculate the mean from each sample and keep it in a list.
                - Calculate the mean of the 400 sample means.
                - Calculated mean of means is off by 39,577.

Lecture 4.3- Mystery Sampling

Comparing Two Samples

- "MysterySample <- c(3706690, 159358, 106405, 55519, 53883)
- mean(MysterySample)"
    - "Is this a sample of U.S. states or is it something else?"
        - "MysterySample is not a sample of states. The mean of MysterySample is just too small to be very likely to be a sample of states."
- **Question: Why is it useful to compare two samples?**
    - When comparing samples, we can see their differences or similarities. But most importantly, we look at their differences and compare it to see in what ways they differ and what we can conclude. For example, we can sample the population of new births from the 1960s and the 2000s and see the causes/factors into why each decade's population differs. We can also compare samples of fan bases of celebrities and see which one is more popular and why.