STAT 230 (Spring 2021): Problem Set 1 (PS1)

Chris Murphy

due by 10:00PM ET on Tuesday, February 23rd

Instruction

I *strongly* encourage that you read the questions as soon as you get the assignment. This will not only help you start thinking how to solve them, but also allow you to have sufficient to get help if you ever need it. In case of questions, or if you get stuck, please don't hesitate to email me (though I would appreciate if you could allow me to have at least 24 hours before the deadline to get back to you). Better yet, you can make a post on our **Q & A Forum** on Moodle, and/or come visit me or TA during our office hours:

- (Shu-Min) Tuesday 3:30 - 5:00PM ET
- (Shu-Min) Thursdays 4:00 - 5:00PM ET
- (Enoch) Sundays 10:00AM - 12:00PM ET

Also, remember that the Statistics & Data Science (SDS) Fellows have Zoom Drop-in hours: Sunday-Thursday evenings from 7:00-8:00PM ET, plus Tuesday & Thursday mornings from 9:00-10:00AM ET (links are available on Moodle). I encourage you to use this resource; the fellows are able to help with questions regarding conceptual understanding of the course material, as well as R and RMarkdown.

Steps to proceed:

1. Download the file *PS1.Rmd* from Moodle
2. Upload the file to the RStudio server (r.amherst.edu)
3. Replace "YOUR NAME HERE" with your name at the top of this document, as well as the *date* part
4. Add in your responses below where it is marked **SOLUTION:**
5. Run **Knit to PDF** under "Knit"
6. Once you are done, upload the knited pdf to **Gradescope**

Problems to turn in: See below

Problem 1:

What is the **sampling distribution** of a sample statistic (like a sample proportion)? Explain what a sampling distribution is, IN YOUR OWN WORDS. (Yes, you can use whatever you learned from our Wednesday class, including the shared Google Docs, but please re-write the statement so that it makes sense to you and *it's yours*.)

SOLUTION: Typically dislayed in a Histogram, the sampling distribution of a sampling statistic is the given distribution of the certain statistic once a simulation has been condcuted on the sample.The sample statistic with vary throughout your simulation and the sample distribution is how we analyze the differences in the sample statstic over the course of the simulation.

Problem 2

[**Modified from IS5-Chapter13, Q13.36**] First USA, a major credit card company, is planning a new offer for their current cardholders. The offer will give double airline miles on purchases for the next 6 months, if the cardholder goes online and registers for the offer. To test the effectiveness of the campaign, First USA recently sent out offers to a random sample of $50,000$ cardholders. Of those, $1,184$ registered. An Amherst student taking STAT135 carefully calculated the corresponding 95% Confidence Interval for the true proportion If those cardholders who will register for the offer, and found it is [0.022, 0.025]. Carefully explain what this interval means (especially what "95% confidence"" means) in a way so that your sibling(s), parents, and/or high-school buddies can understand and write your interpretation below. (You are encouraged to discuss with your Work Team or Study Group to see if others have good ways interpreting this interval too!)

SOLUTION: We can be 95% confident that the true proportion of cardholders who will register for the offer lies between 0.022 and 0.025 of the cardholders. This means our study tells us that we can be 95% confident that that the true percentage of cardholders who will register for the offer is between 2.2% and 2.5% of the cardholders.

Problem 3

[**Modified from IS5-Chapter13, Q15.26**] In 1980, it was generally believed that congenital abnormalities affected about 5% of the nation's children. Some people believe that the increase in the number of chemicals in the environment has led to an *increase* in the incidence of abnormalities. A recent study examined 384 children and found that 46 of them showed signs of an abnormality. Another Amherst student taking STAT135 carefully calculated the corresponding p-value and reported it to be extremely small (1.746768×10^{-10}). Carefully explain what a very small p-value means in the context of this problem, in a way that your sibling(s), parents, and/or high-school buddies can understand and write your interpretation below. To test for their understanding of your interpretation, ask them: "Did this study provide strong evidence that the risk has increased?" and include their responses below as well.

> SOLUTION: The really small p-value tells us that we reject the null hypothesis and that there is strong evidence that the risk for a child abnormalitly has increased. Typically a p-value less than 0.05 will tell us that we reject the null hypotheis.

Problem 4

What is the difference(s) in information provided by using a **confidence interval** versus using a **hypothesis test**? Describe the difference(s) in your own words.

SOLUTION: A 95% confidence interval provides the user with a range where one can be 95 percent confident that the true proportion that they are testing for falls between those intervals. On the other hand a hypothesis test provides the user with a p-value. Depending on if the p-value is above or below 0.05, one can either reject their null hypothesis or fail to reject their null hypothesis.