Statistical Methods in Education Research
Vanderbilt University
Prof. Sean P. Corcoran

Due November 4, 2021
50 total points

Problem Set 8

Instructions: Answer the following questions in their entirety in a separate document. Submit your completed problem set as a PDF document via email to `sean.corcoran@vanderbilt.edu`. Use your last name and problem set number as the filename. Working together is encouraged, but it is expected that all submitted work be that of the individual student.

---

1. You are up late studying, and hungry. You use your last $2 to buy a sandwich out of a vending machine in the basement of Payne Hall. As you unwrap your treat, you notice the inside of the sandwich is a funny color. You ponder whether you should eat the sandwich. In this scenario, what are the Type I and Type II errors associated with your decision to eat the sandwich? (Let H_0 be "the sandwich is safe"). (**3 points**)

2. After recovering from a mysterious sandwich-based food illness, you set out to study whether charter school students perform *better* in mathematics, on average, than traditional public school students. You know that math scores in the population of public school students have a normal distribution with a mean of 420 and standard deviation of 84. You plan to administer the same test to a sample of charter school students to test your hypothesis. Use Stata or a web applet to answer the following questions. (**14 points**)

   (a) What are the null and alternative hypotheses in this problem? (**2 points**)

   (b) Describe what a Type I and Type II error would be in this study. (**2 points**)

   (c) You consider a meaningful difference in test scores to be +21 points. What specific alternative μ would this imply? How large an *effect size* (Cohen's d) would this be? (**4 points**)

   (d) If you were to randomly sample 40 charter school students to take the math test, what is the power associated with your hypothesis test? What is the probability of a Type II error? Use significance level $\alpha = 0.05$. (**3 points**)

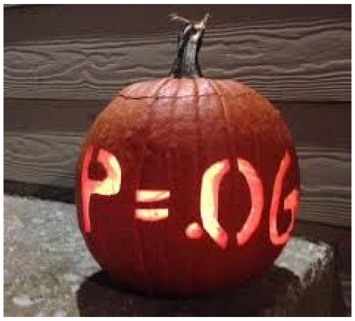
   (e) Holding everything else constant, how large of a sample would you need in order to increase the power to 0.80? (**3 points**)

3. A decision is planned in a test of $H_0 : \mu = 0$ against the alternative $H_a : \mu > 0$, using $\alpha = 0.05$. If the actual μ is 5, the probability of a Type II error (β) is 0.17. (**6 points**)

   (a) Explain the meaning of the last sentence in words. (**2 points**)

   (b) If the test used $\alpha = 0.01$, would the probability of a Type II error be less than, equal to, or greater than 0.17? Explain. (**2 points**)

   (c) If μ were actually 10, would the probability of a Type II error be less than, equal to, or greater than 0.17? Explain. (**2 points**)



4. The Stata syntax below will create a dataset with 200 sample means for $\bar{x}_1$ and $\bar{x}_2$, and 200 sample standard deviations s_1 and s_2. Each sample consists of $n = 50$ random draws from normal distributions in which $x_1 \sim N(0, 5)$ and $x_2 \sim N(2, 5)$. (That is, x_1 has a normal distribution with mean zero and standard deviation 5; x_2 has a normal distribution with mean 2 and standard deviation 5). (**18 points**)

```
set seed #

forvalues j=1/200 {
   clear
   tempfile sample`j'
   set obs 50
   gen x1 = rnormal(0,5)
   gen x2 = rnormal(2,5)
   collapse (mean) x1 x2 (sd) sx1=x1 sx2=x2
   gen n= 50
   save `sample`j''
   }
use `sample1', clear
forvalues j=2/200 {
   append using `sample`j''
   }
```

(a) Run the Stata syntax above, and choose your own starting seed value (#). Take care that your quotations are correct, as they may not copy over correctly from this document into a do-file. Use the `twoway` graphing function with `kdensity` to overlay the resulting distributions of $\bar{x}_1$ and $\bar{x}_2$ (see below). `kdensity` is a kernel density function, a kind of "smoothed" histogram. (**3 points**)

```
graph twoway (kdensity x1) (kdensity x2)
```

(b) Create two new variables that contain the lower and upper bounds of a 95% confidence interval for μ. (Each resulting $\bar{x}_1$—all 200 of them—will have a confidence interval associated with it). Be sure to use the sample standard deviation when creating your confidence interval, not the population standard deviation, and continue to do so in part (c). In what percentage of samples does your confidence interval contain the true population mean $\mu = 0$? In what percentage does it not? (Hint: you can create a third variable that flags whether the CI contains zero or doesn't). (**3 points**)

(c) Create two new variables that equal the t-statistic and p-value from a test of $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. (Each resulting $\bar{x}_1$ will have a t and p-value for this test). Create a histogram to show the resulting distribution of p-values across the 200 random samples. In what percentage of these random samples is $p < 0.05$? Explain in words what this percentage represents. (**4 points**)

(d) You are still interested in the null hypothesis $H_0 : \mu = 0$. However, suppose that—unbeknownst to you—the sample means $\bar{x}_2$ (not the $\bar{x}_1$ that you used in parts b-c) reflect draws from the true distribution of x. That is, x has a mean of 2, not 0. As in part (b), create two new variables that contain the lower and upper bounds of a 95% confidence interval for μ, based on $\bar{x}_2$ (and s_2). In what percentage of random samples does your confidence interval based on $\bar{x}_2$ contain *zero*? Explain in words what this percentage represents. (**4 points**)

(e) Repeat part (c) using $\bar{x}_2$ and continue to test the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. In what percentage of these random samples is $p < 0.05$? In what percentage of these random samples is $p > 0.05$? Explain in words what these percentages represent. (**4 points**)

5. Run the same simulation as in Question #4, but draw `x1` from the exponential distribution, rather than the normal (see syntax below). You can omit the references to x_2 in this problem. Recall from Problem Set 5 that the exponential distribution is heavily skewed, and that the population mean of x_1 is 1/2 here. (**9 points**)

```
gen x1=rexponential(0.5)
```

   (a) Create two new variables that contain the lower and upper bounds of a 95% confidence interval for μ. Be sure to use the sample standard deviation when creating your confidence interval, not the population standard deviation, and continue to do so in part (b). In what percentage of samples does your confidence interval contain the true population mean $\mu = 0.5$? In what percentage does it not? (**3 points**)

   (b) Create two new variables that equal the t-statistic and p-value from a test of $H_0 : \mu = 0.5$ against $H_1 : \mu \neq 0.5$. In what percentage of these random samples is $p < 0.05$? (**3 points**)

   (c) The expected coverage rate of the confidence interval (part a) and the hypothesis test p-values rely on (approximate) normality of the sampling distribution. Your original variable x_1 here was heavily skewed. How did the CI and hypothesis test perform in this case, with the skewed x_1? Explain. (**3 points**)