Statistics 500 - Homework #3, Fall 2020
Due: by noon Friday, 09/11/20

Reading Assignment: Statistical Sleuth, Chapters 3 and 4

1. A major medical center in the Northeastern U.S. conducted a study looking at blood cholesterol levels and incidence of heart attack. Below are data of blood cholesterol levels from 16 people who had a heart attack and 20 people who did not have a heart attack.

(1) Heart Attack				(2) No Heart Attack				
242	186	266	206	182	222	198	192	238
318	294	282	234	198	188	166	204	182
224	276	262	360	178	202	164	230	186
310	280	248	258	162	182	218	170	200

Some summary statistics (all you need if you do the calculations by hand) are
Heart Attack (H): n = 16, sample mean = 265.4, sample s.d. = 43.645
No Heart Attack (N): n = 20, sample mean = 193.1, sample s.d. = 21.623
For all parts of this question, *you can assume that the two groups of people have the same population variance and the cholesterol levels are normally distributed in each group*.

a. Is there sufficient evidence to indicate that the mean cholesterol for people who have had a heart attack is greater than that for people who have not had a heart attack?
(i) Write down the null and alternative hypothesis.
(ii) Find the difference in the two sample means $\bar{Y}_1 - \bar{Y}_2$ and the estimate of the pooled standard deviation S_p.
(iii) Use the values from part (ii) to calculate the test statistic T for the hypothesis test. (Use Hand Calculation.)
(iv) Report degree of freedom for the test statistic T and the p-value.
(v) Give a short (one sentence) conclusion.

b. Using the formula from the lecture notes, calculate a 95% confidence interval for the difference in the two treatment means, and interpret the confidence interval in the context of the problem.

2. In lecture, we derived the variance of the difference in the sample means between two treatments as:

$$Var(\bar{Y}_1 - \bar{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Assuming $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then $Var(\bar{Y}_1 - \bar{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.

a. Compute the value of $Var(\bar{Y}_1 - \bar{Y}_2)$ as a multiple of σ^2 for the sample sizes shown in the following table. The solution is shown for $n_1 = 10$ and $n_2 = 40$.

n_1	n_2	$\text{Var}\left(\bar{Y}_1 - \bar{Y}_2\right)$
1	49	
5	45	
10	40	$0.125\ \sigma^2$
25	25	
30	20	
40	10	
45	5	
49	1	

b. In the table above, the total number of subjects is $n_1 + n_2 = 50$. How would you divide these 50 subjects into two treatments in order to obtain the most precise (least variable) estimate of the difference in the treatment means?

3. Consider an experiment to evaluate whether an herbal extract affects the immune system using cultured cells. The response can be measured using a particular cell assay system. There are two treatments in this experiment: extract and control. Each treatment will be randomly assigned to a tube of cells. A researcher planned an experiment with 6 experimental units (tubes) of each treatment. The variability between experimental units has been measured in other experiments and the sample standard deviation s = 4. We don't know whether the herbal extract will increase or decrease the immune response, a 2-sided test will be used to answer the research question. And we assume equal variance across the two treatments.

(a) What is the standard error of the difference between the two means?
(b) How many degrees of freedom are associated with the standard error (or equivalently, with the T statistic)? What is the 0.975 quantile of the appropriate t distribution?
(c) If the true difference is 5, what is the power for this experiment while type I error rate is controlled at 5% for this two-sample t-test? (Use hand calculation and then you may use SAS to check your answer.)
(d) Suppose that instead of using 6 experimental units per treatment, we would like to design the experiment to get an informative confidence interval for the difference between two treatment means. How many experimental units per treatment (assuming equal sample size for the two treatment groups) should the research have so that the width of a 95% confidence interval for the mean difference is at most 8? (Use hand calculation and then you may use SAS to check your answer.)
(e) Suppose that instead of using 6 experimental units per treatment, we would like to calculate sample size based on power analysis. Still, we have equal sample size for the two treatment groups. Use the power approach to find the required number of experimental units per treatment to provide 90% power for an alpha=5% test, when the true difference is 5. Still assume that the standard deviation is 4. (Use hand calculation and then you may use SAS to check your answer.)
(f) *The subsequent parts will change one piece of the problem at a time and examine the consequences for the sample size. For each of the following, use the same values as in part (e), except for the quantity that is changed.* ***Will the required sample size be bigger or smaller than the sample size you calculate in (e)?*** *You do NOT need to do the calculation or give the exact number of sample size in order to say whether the sample size becomes bigger or smaller.*

    i) If the required power is 95% instead of 90%, is the required number of experimental units bigger or smaller than your answer for (a)? Remember to keep s=4, diff=5, alpha=5%.

ii) If the “important difference” is 3 instead of 5, is the required number of experimental units bigger or smaller than your answer for (a)? Remember to return back to 90% power, alpha = 5%, and s=4.

iii) If standard deviation is 5 instead of 4, is the required number of experimental units bigger or smaller than your answer for (a)? Remember to return back to 90% power, alpha = 5%, and diff=5.

iv) If alpha is 0.5% instead of 5%, is the required number of experimental units bigger or smaller than your answer for (a)? Remember to return back to 90% power, s=4, and diff=5.

4. A researcher wants to conduct a randomized experiment to examine the effect of LSD on concentrations of chemical norepinephrine (NE) in the medulla region of the brain in rats. Half of the rats available for the study will be randomly assigned to a control group and the other half will be exposed to a specific amount of LSD. Information from other studies suggests the mean concentration of NE brain tissue of normal rats, not exposed to LSD, is about 450 ng/gm, and the population standard deviation is about 60 ng/gm. The researcher decided it would important to detect if exposure to the amount of LSD used in the study causes a difference of 25 ng/gm in mean NE concentration. Using a Type I error level of 0.05, the researcher wants a power of 0.85 of detecting an effect of this magnitude.

   (a) Determine how many rats the researcher needs to have in each group. You may use SAS to help answer this question. Hand calculation is NOT required to answer this question.

   (b) Suppose that by switching to a better supplier of rats and improving lab procedures, the researcher could cut the population standard deviation in half, but these improvements would double the cost of each observation. Would these measures be cost effective?

5. If the goal of analysis is to test H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 > \mu_2$. Derive the formula to calculate sample size while controlling type I error rate at α, with effect size of $\delta = \mu_1 - \mu_2 > 0$, desired power of $1-\beta$, and sample size $n_1 = n_2 = n$. (Hint: Slides 170-171) You can give a formula in expression like the one on the bottom of slide 167.