Math-146 - Exam #2 - Practice (F21)

General Information

- You will have 50 minutes to complete Exam #2
- You will need to use a calculator for one question, a short formula sheet will be provided with CLT results
- The exam with contain 3 questions that are each divided into a few parts

Exam topics

1) **Sampling and study design** - populations vs. samples, sampling bias and variability, types of sampling, observational vs. experimental studies, confounding variables, randomization and its impact on confounding variables, blinding, placebos, and other sources of bias
2) **Confidence intervals and bootstrapping** - point vs. interval estimates, margin of error, definition and interpretation of a confidence interval, bootstrapping, bootstrap distributions, the 2-SE method, the percentile bootstrap
3) **Confidence intervals and Central Limit theorem** - How to calculate a P% confidence interval estimate when given the necessary components

Types of content to expect

- 1 question resembling those from our textbook (ie: structured similarly to those in HW #3 and HW #4)
  - Expect 1-2 components related to either sampling or study design
  - Expect 1-2 components that are a review of topics from last time (ie: explanatory and response variables, appropriate types of graphs, etc.)
  - Expect 1-2 components related to confidence intervals and/or bootstrapping
- 1 questions pertaining to concepts, examples, or topics emphasized in our lecture slides and in-class labs
  - Expect 1-2 components related to either sampling or study design
  - Expect 1-2 components that are a review of topics from last time (ie: explanatory and response variables, appropriate types of graphs, etc.)
  - Expect 1-2 components related to confidence intervals and/or bootstrapping

Formulas

You will be given this table on the front page of Exam #2:

Estimate	Standard Error	CLT Conditions
$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$np \geq 10$ and $n(1-p) \geq 10$
$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	normal population or $n \geq 30$
$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i \in \{1, 2\}$
$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	normal populations or $n_1 \geq 30$ and $n_2 \geq 30$
r	$\sqrt{\frac{1-\rho^2}{n-2}}$	normal population (both vars) or $n > 30$

Question #1

A Swedish study investigated the long-term effects of overeating for a short period. The researchers recruited 36 healthy, normal weight individuals and randomly split them into two groups that each contained 18 subjects. Members of the first group were told to increase their calorie intake by 70% while limiting their physical activity for a period of four weeks. Members of the second group were told after an initial weight-in that they did not qualify for the study, but they might be contacted in the future.

Not surprisingly, the weight and body fat of subjects in the first group increased during the study, then began to decrease after the study ended. When the researchers followed-up on each of the 36 subjects 2.5 years later, they found the mean weight gain for participants in the first group was 6.8 lbs with a standard deviation of 5.1 lbs, while the second group (who were told they didn't qualify) had a mean weight gain of 0.6 lbs with a standard deviation 0.9 lbs.

Part A: What are the explanatory and response variables in this study?

Part B: What is the relevant parameter in this study, and what is the point estimate of it? Please define each using proper statistical notation and provide a numerical value for the point estimate.

Part C: Using the results of Central Limit theorem, find the standard error of the point estimate. What conditions must be met for a confidence interval estimate created using this standard error to be valid?

Part D: Using $c = 2.101$, the value from a t-distribution with $df = 17$ needed to achieve 95% confidence, calculate *and interpret* a 95% CI estimate for the difference in weight gain of these two groups.

Part E: Suppose a 99% confidence interval suggests a difference in weight gain of zero is not plausible for the populations represented by these data. Given this result, is it appropriate to conclude causation from this study? Briefly explain (in 1-2 sentences).

Question #2

Between 1972-1974 researchers seeking to better understand the relationship between smoking and all-cause mortality visited the homes of a random sample of adults living in Whickham, a small town in northeast England, collecting data on their smoking behavior. Twenty years later, a follow-up was performed to determine whether each of these subjects were still alive or not. This application involves a subset of the original data containing only women who were either current smokers or had never smoked.

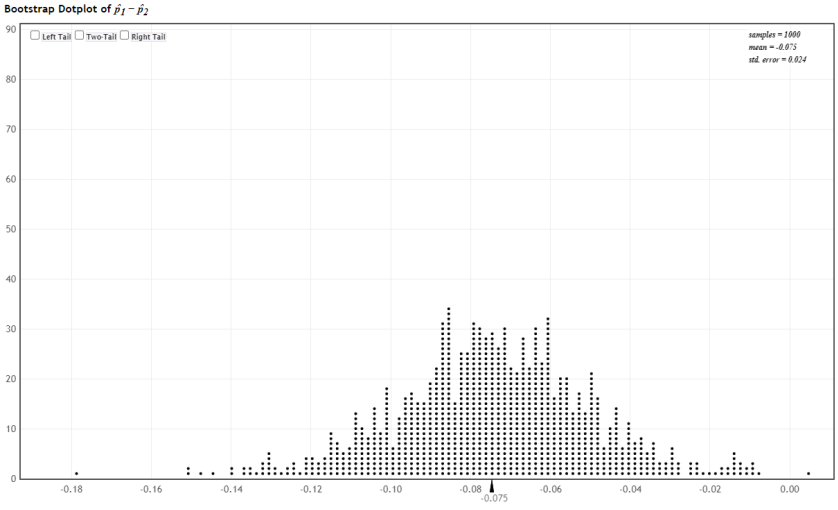
The tables below summarize three variables the researchers collected: age group, smoking status, and survival at the 20-year follow-up

	Alive	Dead	Total
Non-smokers	502	230	732
Smokers	443	139	582
Total	945	369	1314

	Alive	Dead	Total
18-54	765	71	836
55-64	145	91	236
65+	35	207	242
Total	945	369	1314

	Nonsmokers	Smokers	Total
18-54	418	418	836
55-64	121	115	236
65+	193	49	242
Total	732	582	1314

Figure 1 (shown below) displays a bootstrap distribution of the difference in survival proportions for smokers (group 2) and non-smokers (group 1). 1000 different bootstrap samples were created from the original data.



Original Sample

Group	Count	Sample Size	Proportion
Group 1	502	732	0.686
Group 2	443	582	0.761
Group 1-Group 2	59	n/a	-0.075

Bootstrap Sample

Group	Count	Sample Size	Proportion
Group 1	507	732	0.693
Group 2	459	582	0.789
Group 1-Group 2	48	n/a	-0.096

Part A: Consider the explanatory variable "smoking status" and the outcome variable "20-year survival". Is there an association between these variables in these data? Justify your answer using conditional proportions.

Part B: Looking at the bootstrap distribution in Figure 1, briefly explain what each dot in the bootstrap dotplot represents.

Part C: Using the information in Figure 1, use the 2-SE method to find a 95% confidence interval estimate for the difference in survival proportions.

Part D: Based upon your responses to Parts A-D, and considering the other information provided in the prompt for Question #2, characterize each of the following statements as either "true" or "false" (include a brief 1-3 sentence explanation for each).

i) The 95% confidence interval estimate for the difference in survival proportions suggests that we cannot conclude with confidence whether smokers or non-smokers are the group that are more likely to survive in the population represented by these data.

ii) Smokers had a higher rate of survival in the sample data when compared to non-smokers. When taking the 95% CI into consideration, this study provides strong evidence that smoking causes an increase in survival.

iii) The higher survival proportion in the sample of smokers is likely explained by the confounding variable of age. In these data non-smokers tended to be older, and older residents were more likely to die during the 20-year study period.