PH 142 Summer 2021 - Midterm III

The exam is open book. This means you can use electronic or hard copies of all class materials and can use datahub or a local version of R/Rstudio if you wish. You may not use the internet to search for the answers or to inform your answers. Using the internet is strictly prohibited and any evidence of this may result in a 0 on the exam.

While you take the exam, you are prohibited from discussing the test with anyone. If you are taking the test after your classmates, you are also prohibited from talking to them about the test before you take it. Evidence of cheating may result in a 0 on the exam and be reported to the Student Conduct Board.

Berkeley's code of conduct is here: https://sa.berkeley.edu/code-of-conduct. See Section V and Appendix II for information about how UC Berkeley defines academic misconduct. In particular, note the sections on cheating and plagiarism.

UC Berkeley Honor Code
"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." Please carefully read the statements below, and indicate your understanding and intent to adhere to the UC Berkeley Honor code by typing your name in the space below. I agree not to engage in any of the following behaviors:

- Copying or attempting to copy from others during an exam or on an assignment.
- Communicating answers with another person during an exam.
- Pre-programming a calculator or other personal electronic device to contain answers, or using other unauthorized information for exams.
- Using unauthorized materials, i.e. prepared answers.
- Allowing others to do an assignment or a portion of an assignment for you, including the use of a commercial term-paper service.
- Submitting the same assignment for more than one course without prior approval of all the instructors involved.
- Collaborating on an exam or assignment with any other person without prior approval from an instructor.
- Taking an exam for another person or having someone take an exam for you.
- Altering a previously graded exam or assignment for the purpose of a grade appeal or of gaining points in a re-grading process.
- Submitting an electronic file the student knows to be unreadable or corrupted instead of a completed assignment.

Type your name and SID below :

Name:

Enter your name:

Enter your SID:

INSTRUCTIONS:

1.Use Adobe Reader or Acrobat as a stand-alone application (NOT in a browser) to complete this assignment. (this software can be accessed for free for UCB studentshttps://software.berkeley.edu/adobe-creative-cloud)

2.Give your responses ONLY in the space provided. Do NOT add any additional textboxes.

3.Please rename the file LASTNAME_FIRSTNAME_Midterm3_Summer2021.pdf

Unless otherwise specified in the question, format your answers according to the following guidelines:

- present your answers rounded to two decimal places
- present proportions as % values (40.50% rather than .405)

** MAKE SURE YOU ARE WORKING WITH THIS DOCUMENT IN ADOBE AND YOU ARE NOT IN A BROWSER WINDOW **

Question 1 [5 pts total]

Use the following excerpt to answer the questions below.

Black individuals have historically received a lower average annual salary compared to individuals of other races/ethnicities. With the Black Lives Matter Movement educating the public about racial inequities that black individuals experience, sociologists wondered whether the average annual salary for Black individuals was significantly different in 2020 compared to the previous year. To investigate this, a group of sociologists collected a random sample of 100 black individuals from 25 companies and obtained their average 2019 and 2020 salaries. They found that the average salary for all 100 individuals was $42,460 in 2019, $44,220 in 2020, the standard deviation of the difference in salaries was $1,600, and the standard deviation of salaries was $6,450 in 2019 and $5,275 in 2020.

a. [1 pt] What is the outcome of interest and what type of variable (nominal, ordinal, continuous, discrete) is it?

b. [1 pt] What test statistic should you use to investigate the difference in salaries?

c. [1 pt] What is the corresponding null hypothesis in the context of this problem? Is the alternative hypothesis one-sided or two-sided?

Null Hypothesis:

One- or Two-sided:

d. [2 pts] If you wanted to estimate the difference in mean salaries within ± \$200 with 95% confidence, how many individuals would you need to sample? The critical value is 1.984. Round up to the nearest whole number.

Question 2 [5 pts total]

Use the following excerpt to answer the questions below.

In the state of California, there is an average 7,500 individuals enrolled in Medicare Insurance per 10,000 eligible individuals with a known standard deviation of 1,000 enrollees per 10,000 eligible. This average was calculated from insurance data collected from all cities in California and appears to be normally distributed. You are interested in whether the average number of those enrolled in Medicare per 10,000 eligible in the Bay Area is greater than the state of California's average; they find that $\bar{x}$ = 7,800 Medicare enrollees from a random sample of 100 eligible Bay Area individuals.

a. [1 pt] State the null and alternative hypothesis using statistical notation.

H_0:

H_A:

b. [2 pts] Use the information above to decide the appropriate test statistic and calculate the test statistic value by hand. Show your work.

Test Statistic:

Work:

Value:

c. [1 pt] Next, you want to calculate the power of the test. Write the two lines of code to calculate the power assuming alpha = 0.05, assigning the first value to an object called `critical_val`.

Line 1:

Line 2:

d. [1 pt] You calculate a power of 91.2%. Interpret this value in the context of the problem.

e. [1 pt] **Extra Credit** What would you expect to happen to the power if you increased your sample to 200 individuals? Why?

Question 3 [5 pts total]

Non-parametric testing w/ COVID Antibody Data

A study was conducted to determine the effectiveness of the BNT162b2 mRNA COVID-19 vaccine by Jalkanen et al. (https://doi.org/10.1038/s41467-021-24285-4). One indicator of effectiveness is assessing the concentration of the anti-S1 IgG antibody of one group of volunteers before and another group of volunteers after vaccination. The researchers want to see if there is any difference in IgG antibodies between the two samples. The following table is a subset of anti-S1 IgG concentrations at at 0 days and 42 days after vaccination:

0 days	Rank	42 days	Rank
30	a	111	b
26	6	112	12
23	4	107	8
21	3	111	d
19	c	127	14
17	1	108	9
25	5	115	13
Sum	e	Sum	f

a. [2 pts] Fill in the table with the ranks for each missing data point and the total sum for each sample.

a:
b:
c:
d:
e:
f:

b. [1 pt] Calculate the μ_w and σ_w statistics. Provide the mean in its entirety and round the standard deviation to the sixth decimal place. Only provide the answer within the boxes.

μ_w:
σ_w:

c. [1 pt] Calculate the Z_w statistic. Round your answer to two decimal places.

Z_w:

d. [1 Point] Why is this test an appropriate test for these data and the comparison we are making?

Z_w

Question 4 [4 pts total]

As covered in lab, the National Health and Nutrition Examination Survey (NHANES) dataset covers a wide range of health data from randomly chosen households in the United States.

Suppose a graduate student at Berkeley's School of Public Health wants to determine whether or not there's a difference between the proportions of those who have health insurance among those with and without asthma.

a. [1 pt] State the null and alternative hypotheses in the context of the question the graduate student wants to answer.

H_0:

H_a:

Filtering the NHANES dataset by the above restrictions resulted in the following values:

Number of Respondents with Asthma	2135
Number of Respondents without Asthma	366
Number of Respondents with Asthma and Health Insurance	1673
Number of Respondents without Asthma and with Health Insurance	305

b. [1 pt] Find the following proportions and round them to the nearest fourth decimal place.

Proportion of those with health insurance with asthma ($\hat{p}_1$):

$\hat{p}_1$:

Proportion of those with health insurance without asthma ($\hat{p}_2$):

$\hat{p}_2$:

c. [1 pt] Provide R code to calculate the p-value using the Wilson Score Method. Assign your test to the variable `asthma_prop_test`.

d. [1 pt] Calculate the p-value (report this to 2 decimal places) and interpret the results of this test in your own words.

Question 5 [7 pts total]

You are working in the Department of Agriculture and are given data about the effects of 3 different types of fertilizer on crop yield. You are tasked with analyzing whether any of the fertilizers produce a significantly better average crop yield. The dataset is called `crop_data` and the variables of interest are `crop_yield` (number of crops/ounce of fertilizer) and `fertilizer` (type of fertilizer).

a. [2 pts] Write the 2 lines of code that give you the output below, assigning your first line of code to an object called `crop`. What can you conclude about the differences between the fertilizers and their crop yields based on these results?

```
## # A tibble: 2 x 6
##   term          df sumsq meansq statistic   p.value
##   <chr>      <dbl> <dbl>  <dbl>     <dbl>     <dbl>
## 1 fertilizer     2  6.07  3.03       7.86    0.000701
## 2 Residuals     93 35.9   0.386     NA    NA
```

Code:

Conclusions:

b. [2 pt] What statistical method can you use to determine *which* fertilizer is different and why is this method better than conducting many pairwise tests? (hint: this can be easily done by adding a short statement to the above code)

c. [1 pt] After using the method from part b, you find that fertilizer a, labeled `fert_a` in dataset, has a significantly different crop yield than the other fertilizers. You subset your data into `crop_subset` to include the variable `fert_a` which is the amount of fertilizer a (in ounces) and the corresponding `crop_yield`. You want to test whether there is a linear relationship between these 2 variables. Write the null and alternative hypotheses in the context of this problem.

H_0:

H_a:

d. [2 pts] You use R to fit a linear model of your data and assign it to an object called `fit` and find an intercept of 3.72 ounces and a slope of 1.696. You are given another dataframe, `newdata = data.frame(fert_a = 150)`, to predict the crop yield for a `fert_a` value of 150 ounces. Calculate this crop yield by hand using your linear model and write the line of code that outputs the prediction interval.

Prediction:

Code:

Question 6 [4 pts total]

Researchers are interested in determining whether a new antidepressant is significantly associated with an improvement in depressive symptoms. They conduct an observational study with 200 patients from a mental health clinic: 110 who take the new antidepressant and 90 who do not. The table below displays data on their symptom improvement status.

	Improvement	No Improvement
Antidepressant	64	46
No Antidepressant	36	54

a. [1 pt] What statistical test would you use to determine whether there is an association between taking the new antidepressant and observing improvement in depressive symptoms? State the null hypothesis.

Statistical test:

H_0:

b. [1 pt] Calculate the expected values for each of A through D by hand. Show your work.

	Improvement	No Improvement
Antidepressant	A	B
No Antidepressant	C	D

A:

B:

C:

D:

Work:

c. [2 pts] Use the expected values you calculated above to calculate the test statistic by hand. Show your work.

Work:

Test statistic:

Exam feedback:

If you experienced any issues with your exam please describe them here: