PH 142 Summer 2021 - Midterm I

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 [1 point]:

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 b e 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_Midterm1_Spring2021.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 [6 pts total]

Because county hospitals often have fewer available resources compared to private hospitals, researchers were interested in studying the relationship between treatment success and type of hospital. The following data looks at the relationship between the type of hospital that patients go to (Private, County) and the outcome of the treatment they receive, which can be a success or failure. A third variable is the severity of the illness patients have when they go to the hospital (severe or not severe).

a. [1 pt] Fill in the blanks for the following two-way table:

	Success	Failure	Total
Private	50	A	100
County	B	32	100
Total	C	82	D

A: ☐

B: ☐

C: ☐

D: ☐

b. [1 pt] What are the marginal distributions of successful and unsuccessful treatments in this population? Use percentages and round to two decimal places.

Successful: ☐

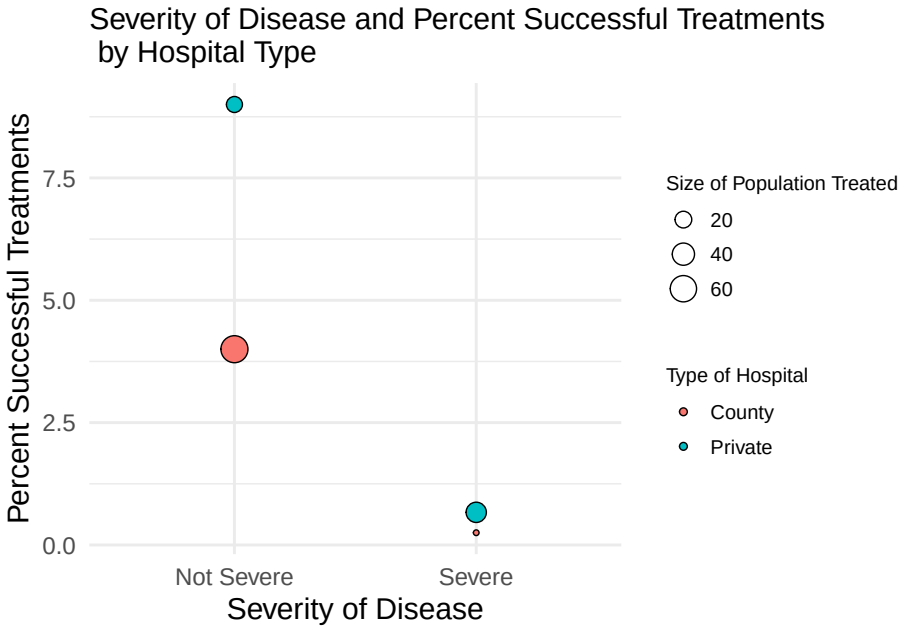
Unsuccessful: ☐

c. [2 pts] What is the distribution of successful treatments conditional on the type of hospital?

Successful | Private: ☐

Successful | County: ☐

d. [2 pts] Based on the visualization below, it is evident that within the levels of severity, there is a higher rate of success for private hospitals compared to county hospitals. In 1-3 BRIEF sentences, identify the cause of this phenomenon, and explain why that is the cause.



Question 2 [7 pts total]

Below is a dataframe called OFCdata that is from a neuroscience research study that examines the relationship between two regions in the orbitofrontal cortex of the brain (OFC1 and OFC2). The researcher plant electrodes into both brain regions for several individuals and records the activity level during different independent stimuli. The results are shown below.

```
## # A tibble: 6 x 2
##     OFC1  OFC2
##    <dbl> <dbl>
## 1  0.759  6.31
## 2  4.10   7.94
## 3  2.01   9.46
## 4 12.6   11.0
## 5  4.98  11.1
## 6  1.92   8.90
```

a. [2 pts] You decide to fit a linear regression to the data. Fill in the blanks to generate the output shown below.

```
fit <- __[A]__( __[B]__ ~  __[C]__, OFCdata)

library(broom)

__[D]__(fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    6.18     0.145       42.6 7.26e-129
## 2 OFC1           0.469    0.0220      21.4 4.90e- 62
```

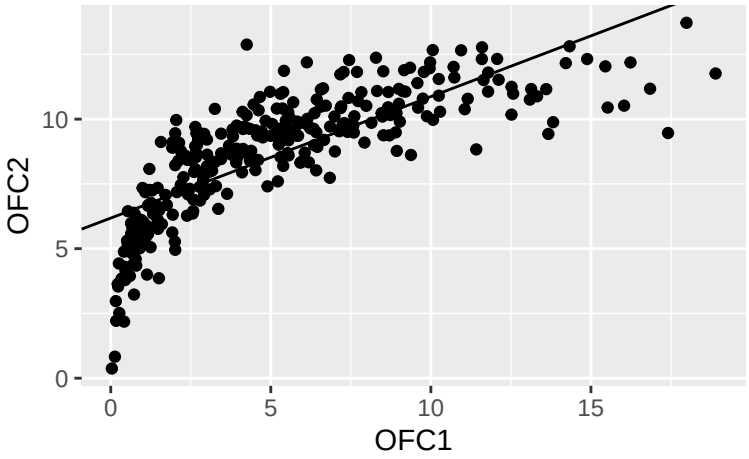
A: [ ]

B: [ ]

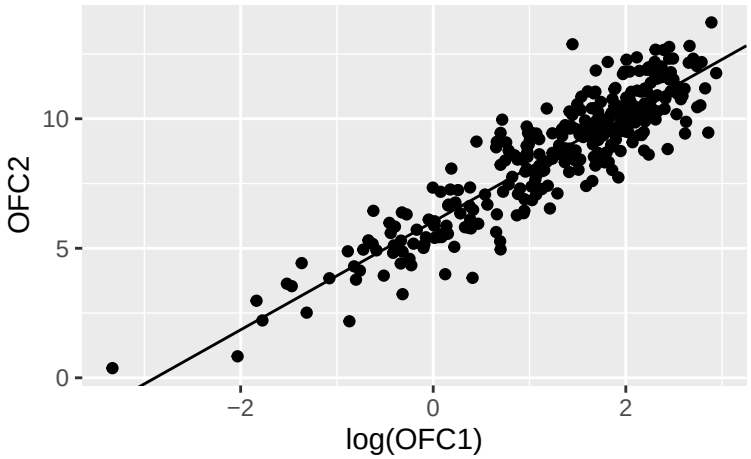
C: [ ]

D: [ ]

b. [1 pt] Below is a plot of the OFCdata with the line of best fit that you generated in part a. Based on this plot, visually describe the relationship between OCF1 and OCF2.



c. [2 pts] You then perform a logarithmic transformation on the OFC1 data points and generate a new model. Without being given any numbers, which plot (original vs. transformed) has a stronger correlation coefficient? Why?



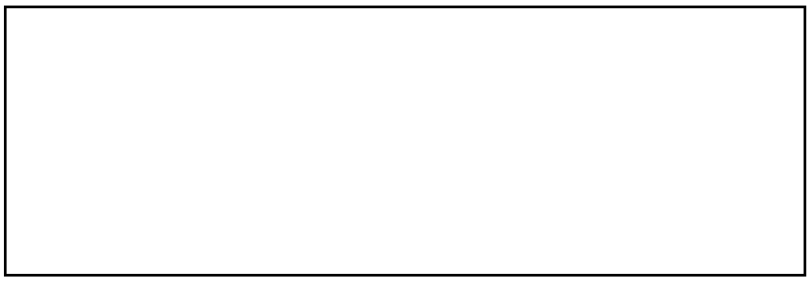
d. [1 pt] Which of the following is a plausible value for the correlation coefficient between log(OFC1) and OFC2?

- 0.6
- 0.9
- 1.0
- 0.3

r =

e. [1 pt] Interpret the slope parameter from the output below in the context of the problem.

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     6.03    0.0937      64.4 7.73e-177
## 2 log(OFC1)       2.09    0.0574      36.4 1.22e-111
```



EXTRA CREDIT [1 pt] You are given one more observation of an individual with an OFC1 value of 40.14 and would like to use your transformed model to predict their OFC2 value. Calculate the predicted OFC2 value and round to two decimal places. Is this prediction appropriate given your data? Why or why not?

Question 3 [6 pts total]

Please refer to the following passage for question 3.
The following is an excerpt from Toifail et. al The Lancet 2018:

Background: Poor nutrition and hygiene make children vulnerable to delays in growth and development. We aimed to assess the effects of water quality, sanitation, handwashing, and nutritional interventions individually or in combination on the cognitive, motor, and language development of children in rural Bangladesh.

Methods: In this cluster-randomized controlled trial, we enrolled pregnant women in their first or second trimester from rural villages of Gazipur, Kishoreganj, Mymensingh, and Tangail districts of central Bangladesh, with an average of eight women per cluster... Groups of eight geographically adjacent clusters were block-randomised into six intervention groups and two control groups.

The six intervention groups were: chlorinated drinking water; improved sanitation; handwashing with soap; combined water, sanitation, and handwashing (WASH); improved nutrition through counselling and provision of lipid-based nutrient supplements; and combined water, sanitation, handwashing, and nutrition (WASH+N). Here, we report on the prespecified secondary child development outcomes: gross motor milestone achievement assessed with the WHO module at age 1 year, and communication/gross motor/personal social/combined scores measured by the EASQ at age 2 years. Masking of participants was not possible.

Findings: At year 1, compared with the control group, the combined water, sanitation, handwashing, and nutrition group had a higher rate of attaining the standing alone milestone (hazard ratio $1 \cdot 19$, 95% CI $1 \cdot 01$–$1 \cdot 40$), and the nutrition group had a higher rate of attaining the walking alone milestone ($1 \cdot 32$, 95% CI $1 \cdot 07$–$1 \cdot 62$). The combined water, sanitation, handwashing, and nutrition group had a higher rate of attaining the walking alone milestone than those in the water, sanitation, and handwashing group ($1 \cdot 29$, $1 \cdot 01$–$1 \cdot 65$). At 2 years, we noted beneficial effects in the combined EASQ score in all intervention groups, with effect sizes smallest in the water treatment group (difference $0 \cdot 15$, 95% CI $0 \cdot 04$ to $0 \cdot 26$ vs control) and largest in the combined water, sanitation, handwashing, and nutrition treatment group ($0 \cdot 37$, $0 \cdot 27$–$0 \cdot 46$).

a. [1 pt] What type of a problem is this study interested in addressing? (Options: Descriptive, Prediction, Causative/Etiologic)

b. [1 pt] This is a(n) _____ study design. (Options: Observational, Experimental, Case-Series, Cross-Sectional)

c. [1 pt] What type of variable is the exposure of interest? (Options: Ordinal, Nominal, Continuous, Discrete)

d. [1 pt] Who is the target population in this study? (1 sentence)

e. [1 pt] To what population(s) might researchers want to generalize these results?

f. [1 pt] In order for researchers to generalize these results, the study needs to be _____ valid.

Question 4 [3 pts total]

Please refer to the following passage for question 4. All children (under 19 years of age) diagnosed with neuroblastoma over the period May 1, 1992, to April 30, 1994, were contacted for parent interviews in a study conducted by Olshan et. al 92002). Interviews focused on parental occupational history, medication use, and pregnancy histories, with a particular interest in maternal vitamin use. Similar interviews were also conducted to parents of healthy children who were selected by random digit-dialing.

a. [1 pt] Is this study an observational or experimental study?

b. [2 pts] Is this study conditional on the exposure or the outcome? Based on your answer, what type of study design is this?

Conditional on:

Study design:

Question 5 [5 pts total]

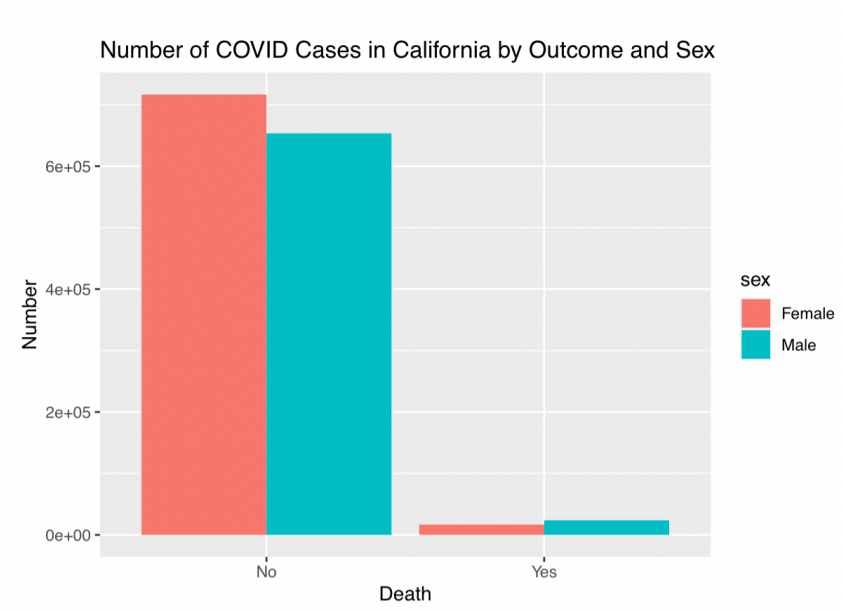
You are a biostatistician working for the California Department of Public Health. You are given a dataset (called `COVID_CA`) with COVID-19 mortality rates in California, Oregon, and Washington and are told to clean, manipulate, and analyze the data to report back to your team.

a. [1 pt] You first want to subset your data to only contain observations from California and assign it to a new dataset called COVID_CA. The variable 'res_state' describes which state each individual resides in. Write the appropriate line of code that outputs this information.

b. [1 pt] Next, you want to see how many observations are in your COVID_CA dataset. Write the appropriate line of code that outputs this information.

c. [2 pts] You'd like to stratify your COVID_CA data to visualize the mortality rates by sex. You create the 2x2 table called COVID_CA_2 shown below with the variables death and sex. Fill in the code that will output the plot below.

```
## # A tibble: 4 x 3
##   sex    death number
##   <chr>  <chr>  <dbl>
## 1 Female No    716595
## 2 Female Yes    16724
## 3 Male   No    653407
## 4 Male   Yes    23784
```




```
plot <- ggplot(COVID_CA_2, aes(x = ___[A]___, y = number)) +
geom_bar(aes(___[B]___ = ___[C]___), stat = ___[D]___, position = "dodge")

plot
```

A:

B:

C:

D:

d. [1 pt] What conclusions about sex and COVID mortality in California can you make based on the visualization shown above?

Question 6 [4 pts total]

In a study conducted by The Diabetes Prevention Program Research Group, 3,234 participants from 27 clinics in the U.S. were enrolled in the Diabetes Prevention Program (DPP) from 1996 to 1999. The analysis included 2,155 participants randomly assigned to the metformin (1,073) or placebo (1,082). Below is a simulated random sample of the weight loss (in pounds) recorded from those assigned to the metformin group.

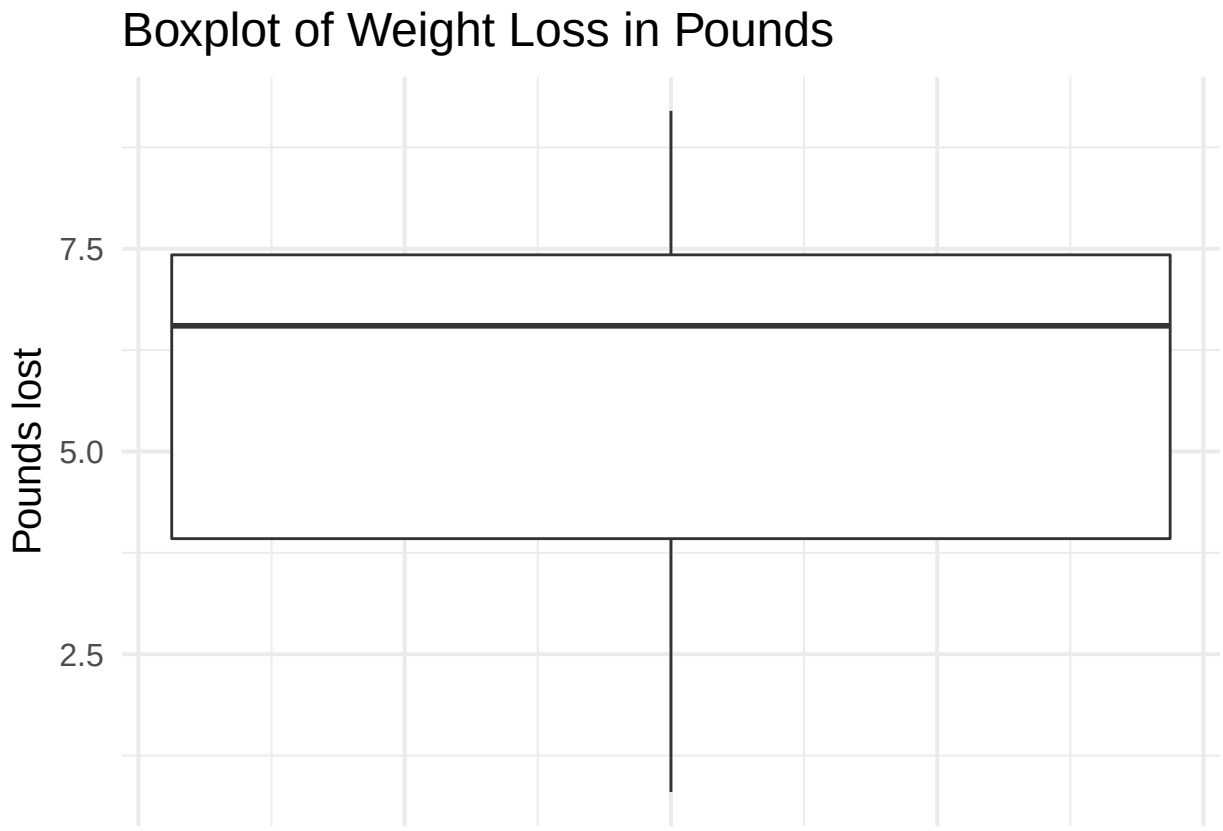
```
##    participants weight_loss
## 1             1         0.8
## 2             2         1.3
## 3             3         2.4
## 4             4         3.4
## 5             5         5.5
## 6             6         6.0
## 7             7         6.3
## 8             8         6.8
## 9             9         6.9
## 10           10         7.2
## 11           11         7.5
## 12           12         8.0
## 13           13         8.5
## 14           14         9.2
```

a. [1 pt] Fill in the line of code to generate the boxplot of the weightloss_data shown below.

```
ggplot(data = weightloss_data, aes(__[A]__ = __[B]__)) + geom_boxplot()
```

A: [ ]

B: [ ]




```
##   min    Q1 median    Q3 max
## 1 0.8 3.925   6.55 7.425 9.2
```

b. [2 pts] Describe the distribution of the weight_loss variable as completely as possible based on the boxplot from part a.

Exam feedback:

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