Fall 2022 Midterm I V2

The exam is closed book and closed notes. You are allotted one double sided "cheat sheet" which may contain typed or handwritten notes. You may also use a calculator. Your phone is not allowed as a calculator. Using any resources outside of the aforementioned items is strictly prohibited.

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.

Write your name and SID below.

Enter your name: □

Enter your SID: □

INSTRUCTIONS:

Hand write your responses using a pencil or pen in the space provided. Give your responses ONLY in the space provided. Do not write your responses outside of the provided text boxes. The additional space provided, including space on the back side of the exam can be used as scratch paper but will not be graded.

Phones should be turned OFF proir to the start of the exam and secured in your backpack or another secure location. Do not leave your phone or other electronic devices out. If you need to leave the room for any reason during the exam please flag a GSI and let them know prior to exiting the room. Time will still accrue when you leave the room.

The length of the midterm if 50 minutes. If you finish early and are satisfied with your work you may leave early. Hand in your exam to a GSI, who will verify that they received it.

- 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)
- All logs are natural log base e

Exam Format:

Short Format Questions: 1a, 1b, 2 [3 points]
Quick Response Grouped: 3a, 3b, 3c, 3d [2 points]
Short Format Questions pt. 2: 4, 5, 6a, 6b [5 points]
Long Format Question on Mosquito Control: 7a-f [10 points]
Long Format Question on T2DM in Central Virginia: 8a-f [10 points + 1 bonus]
Optional Feedback Question: 9

1a. [1 point] You work for a healthcare company and your boss wants you to analyze what factors are correlated with a specific genetic disorder. You find that a specific gene, Gene Grey, is associated with your study's genetic disorder. You hope to use this information to predict future findings across genes with similar genetic compositions This is an example of what kind of study:

□ A. Descriptive
□ B. Predictive
□ C. Causitive/Etiologic

1b. [1 point] The association between the genetic disorder and Gene Grey is very strong. Does this imply a causative relationship?

□ A. No, this does not imply a causative relationship
□ B. Yes, this does imply a causative relationship

```
#Answer 2A: B, Predictive
#Answer 2B: A, no this does not imply a causative relationship.
```

2.[1 point] You are given a dataset, `monkeybox` which has 6 columns `id, county, state, num_deaths, population and num_uninsured`. Running the following code `monkeybox %>% select(-num_uninsured)`, the output will have exactly 1 column.

□ True
□ False

```
#Answer: False. Monkeybox will have 5 columns after running the code.
```

3. For each of the following, choose the most appropriate function to visualize the dataset:

3a. [1/2 point] The Diversity, Respect, Equity, Action, Multiculturalism (DREAM) office at UC Berkeley's School of Public Health is interested in creating a visualization of distribution of the number of students in each graduate program (Biostats, Health Policy, etc.).

□ A. `geom_point()`
□ B. `geom_histogram()`
□ C. `geom_line()`
□ D. `geom_bar()`
□ E. `geom_abline()`

3b. [1/2 point]A doctor at Kaiser wants to see the distribution of patient's weight in their clinic.

□ A. `geom_point()`
□ B. `geom_histogram()`
□ C. `geom_line()`
□ D. `geom_bar()`
□ E. `geom_abline()`

3c. [1/2 point] We want to visualize the relationship between miles walked per day and average minutes of sleep per day among Berkeley Public Health students.

□ A. `geom_point()`
□ B. `geom_histogram()`
□ C. `geom_line()`
□ D. `geom_bar()`
□ E. `geom_abline()`

3d. [1/2 point]The Tang Center has student survey data on students level of happiness measured on a scale from zero to one hundred and is interested in creating a visualization of the distribution of students age by happiness level.

□ A. `geom_point()`
□ B. `geom_histogram()`
□ C. `geom_line()`
□ D. `geom_bar()`
□ E. `geom_abline()`

```
#Answer: 3A - D. geom_bar
#        3B - B. geom_histogram
#        3C - A. geom_point
#        3D - B. geom_histogram
```

4. [1 point] An study looking at the causal relationship between environmental pollution and asthma. A student researcher brings up the fact that economic status is associated with both exposure to environmental pollution and asthma. This type of variable is known as a(n) ______________________________ variable.

□ A. confounding
□ B. observational
□ C. supplemental
□ D. experimental

```
#Solution: A confounding
```

5. [1 point]In a randomized control trial both patients and site administrators (including clinicians) are unaware of an individuals treatment status. This is an example of:

□ A. Confounding
□ B. Unethical Behavior
□ C. Blinding
□ D. An impossible scenario

```
#Answer: c. Blinding
```

6. Discharge rates are one metric cost savings in a hospital setting. When trying to compare two quality metrics, statisticians looked at patient emergency room discharge rates for patients with self reported mild, moderate, and severe pain at Kaiser Walnut Creek.

6a. [2 points]Fill in the blanks in the following two-way table.

	discharged	undischarged	Total
mild	340	A	415
moderate	315	B	C
severe	D	80	E
Total	740	260	1000

A:

B:

C:

D:

E:

```
# SOLUTION: A = 75, B = 105, C = 420, D = 185, E= 165
```

6b. [1 point] What is the conditional distribution of discharge status among those who were admitted with severe symptoms? Report your answer as a fraction.

```
# SOLUTION:
# The discharge rate among people with severe symptoms is 51.5% (85/165)
# The not discharged rate among people with severe symptoms is 48.5% (80/165)
```

Question 7[10 points] A mosquito control district wants to do a field experiment to look at the distribution of mosquitoes and standing water sources. Entomologists placed 250 mosquito trap sites and counter the number of standing water sources within a 3-mile radius of the trap sites. They also counted the average number of mosquitoes collected in each trap site over a month. They hypothesize that more standing water allows for a greater number of mosquitoes to lay eggs and have a larger population.

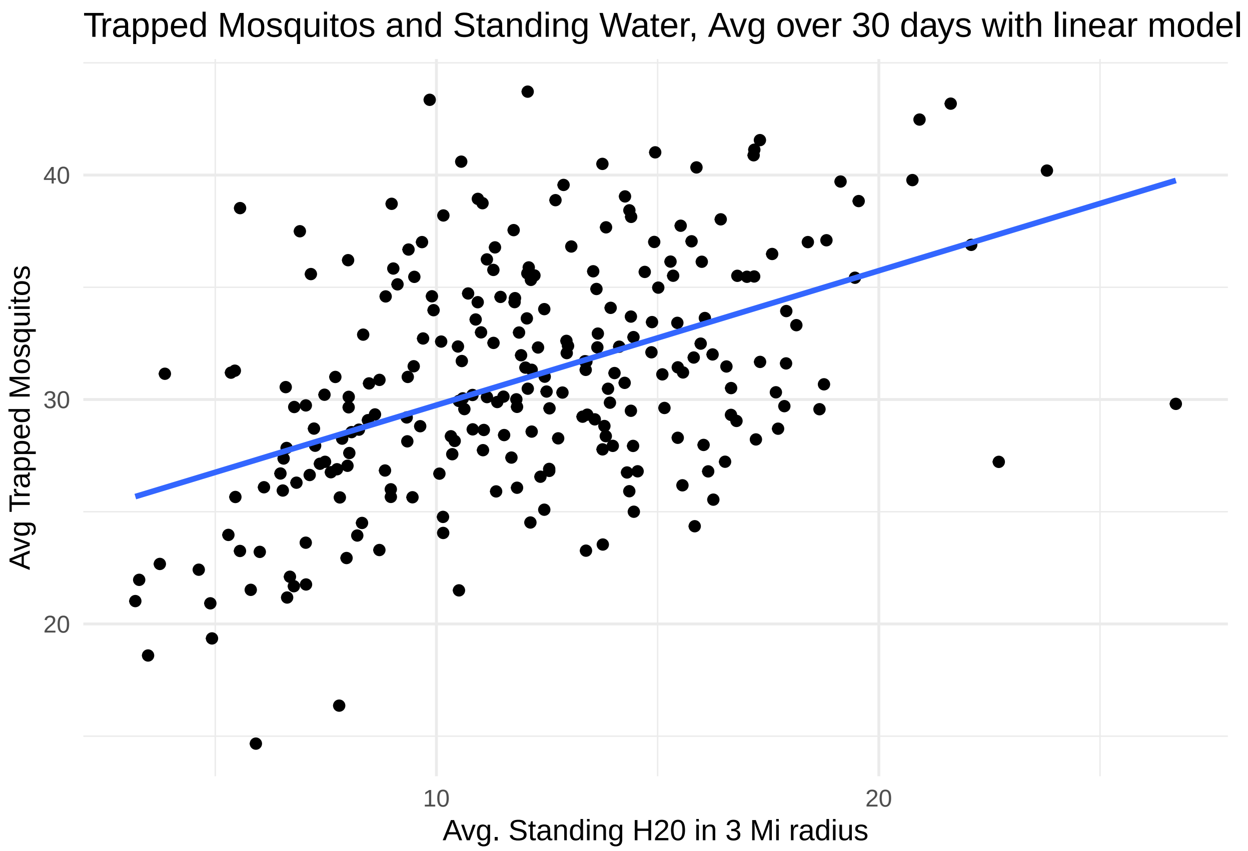
This dataset has two columns: `standing_water`, which is an average count of sources of standing water within a 3-mile radius as observed by a field consultant, and `mosquito`, an average count of the number of trapped mosquitoes as measured by the same field across a calendar month.

7a. [1 point] Given the study description above, which of the following variables are explanatory and which of the variables represents the outcome?

☐ `mosquito` is the explanatory variable and `standing_water` is the outcome variable.
☐ `standing_water` is the explanatory variable and there is no outcome variable.
☐ `standing_water` is the explanatory variable and there is no outcome variable.
☐ Neither `standing_water` nor `mosquito` are related to one another according to the study description.

```
#Answer: C. standing water is the explanatory variable and mosquito is the outcome variable.
```

Trapped Mosquitos and Standing Water, Avg over 30 days with linear model



```
mod1 <- lm(mosquito ~ standing_water, df_mosquito)
tidy(mod1)
```

```
## # A tibble: 2 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      23.8      0.906      26.2  1.97e-73
## 2 standing_water    0.599    0.0716      8.36 4.53e-15
```

```
glance(mod1)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.220         0.217  4.65      69.9 4.53e-15     1  -738. 1482. 1493.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
coorelation_coef <- sqrt(.220)
coorelation_coef
```

```
## [1] 0.4690416
```

You generate a scatter plot of relationship bewtween `standing_water` and `mosquito` and also construct a linear model to study their relationship. The results of the model, named `mod1` are shown above, using the `tidy()` function to generate a table of the results, the `glance()` function to show the R^2 value, and a calculation of the R value as the 3rd output.

7b. [2 points] Using this information, describe the relationship between the two variables. Make sure to include a description of both the generated plot AND the linear model. The final value is the pearson correlation coefficient.

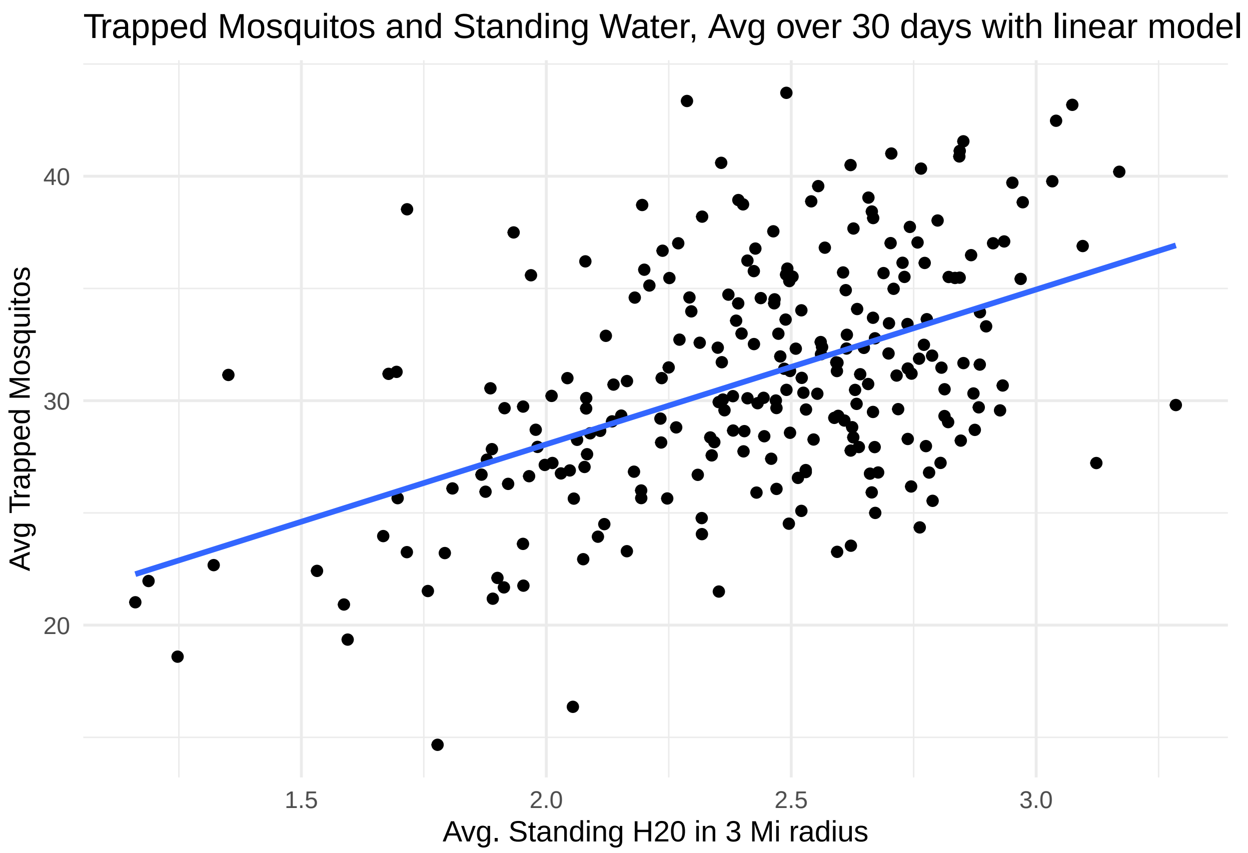
```
#Answer: The points show a lot of variation around the line of best fit.
#The r squared and pearson correlation coefficient show very weak correlation.
#Perhaps there is a better fitting model based on the data
```

7c. [1 point] The R^2 value is .220 Select ALL that are true:

- □ A. The value of R^2 can range from 0 to 1
- □ B. R^2 tells us the value of `mosquito` when `standing_water` is zero
- □ C. The value of R^2 can range from -1 to 1
- □ D. 22% of the variation in mosquito popultions can be explained by the number of standing water sources.
- □ E. R^2 is not valuable when interpreting the model

```
#Answer: A, and D.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
tidy(mod2)
```

```
## # A tibble: 2 x 5
##   term                estimate std.error statistic  p.value
##   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)            14.3      1.86       7.65 4.44e-13
## 2 log(standing_water)     6.90     0.762      9.05 4.37e-17
```

```r
glance(mod2)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.248         0.245  4.57      81.9 4.37e-17     1  -733. 1473. 1483.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```r
coorelation_coef2 <- sqrt(.248)
coorelation_coef2
```

```
## [1] 0.497996
```

The data was transformed to use the log of `standing_water` instead of `standing_water`. The results of the model, named `mod2` are shown above, using the `tidy()` function to generate a table of the results, the `glance()` function to show the R^2 value, and a calculation of the R value as the 3rd output.

7d. [2 points]Provide an interpretation of the slope coefficient generated by `mod2` in the context of this study.

```
#Answer: For every 1 unit change in the log average standing water sources, mosquito counts increased b
```

7e. [2 points] Is the interpretation of the intercept meaningful in `mod2`? Why or why not?

```
#Answer: The intercept is meaningful in this case, since it measures the amount of mosquito we can expe
#Alternative: The intercept might be beyond the range of our data, so we might be extrapolating beyond
```

7f. [2 points] As a next step, the team plans to see if distribution of chlorination tablets has an effect on the mosquito population. The team plans to provide tablets to local community centers in the area and allow residents to pick up tablets. In the context of a trial, how might you critique this approach? Propose an alternative based on your critiques. If you believe this approach works fine, then simply state that you would not make any changes.

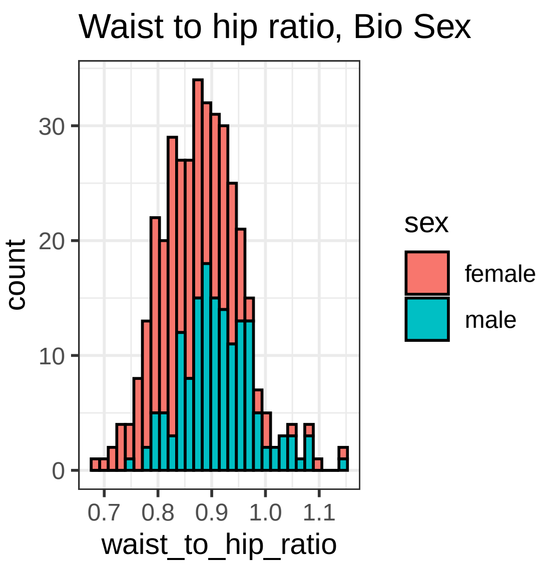
```
#Answer: Student can make the case that this is non-random sample -- convenience sample or volunteerism
#For part 2, student can mention that they would randomize residents who received the tablets, which wo
```

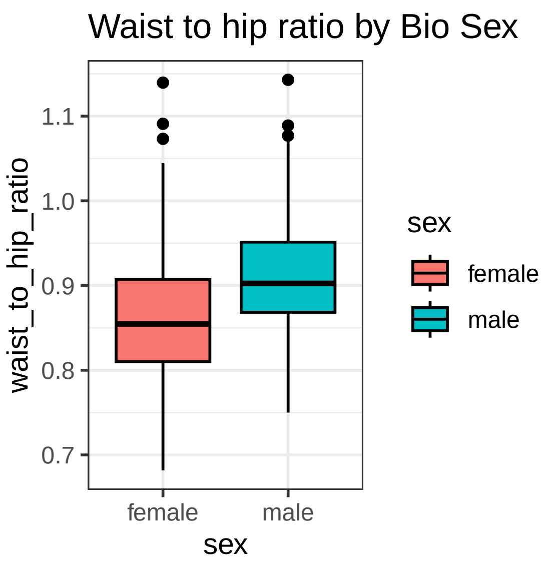
8. [10 points + 1 bonus] Diabetes is a major health issue in the US. In 2018, 34.2 million Americans, or 10.5% of the population, had diabetes. The object `diabetes` contains cardiovascular data that was taken on a random sample of 375 patients in Central Virginia to study the prevalence of cardiovascular risk factors in this population. We are interested in exploring indicators of Type II Diabetes within this sample.

Variable	Description
id	Patient identifier
chol	Cholesterol
stab.glu	Stabilized glucose levels
hdl	High density lipoprotein
ratio	chol/hdl
glyhb	Glycosylated hemoglobin
location	Location
age	Age
sex	Sex assigned at birth
height	is in inches
weight	is in inches
bp.1s	first systolic blood pressure
bp.1d	first diastolic blood pressure
waist	given in inches
hip	circumference given in inches
frame	complexity
time.ppn	minutes after eating that their glucose levels were measured (postprandial time)

8a. [2 points] Waist to hip ratio is a common cited risk factor for diabetes. Write a line of code that creates a new variable, called `waist_to_hip_ratio` that calculates this value, the ratio of waist size over hip size, and adds it to a dataframe called `diabetes_modded`. The original dataframe, for reference, is named `diabetes`, and the variables are: `waist` and `hip` respectively.

```
#Answer: diabetes_modded <- diabetes %>% mutate(waist_to_hip_ratio = waist/hip)
```





8b. [3 points]Using the following plots and in no more than 4 sentences, describe the distribution of waist to hip ratios among the sample population. Be sure to include approximate numerical summaries and remark on differences beteween listed genders, if there are any.

```
#Answer: Students should discuss the following aspects:
#spread: range, IQR
#skew: central tendency, mean, median
#centrality: unimodal
#outliers: present
#They may also discuss the apparent difference between represented genders.
```

8c. [1 point]Suppose you wanted to look at the numerical differences in waist to hip ratio between males and females rather than graphically interpreting it with the plots shown above. Which two dplyr commands would you use to complete this task?

☐ A. filter and arrange
☐ B. rename and select
☐ C. group_by and summarize
☐ D. mutate and sort

```
#Answer - C. group_by and summarize
```

8d. [1 point]This study was conducted in Central Virginia. Researchers might wonder if this study might also say something about a population of individuals who live in San Antonio, Texas. Select the phrase that best expresses the thought that this population may be representative of the population in San Antonio.

☐ A. External Validity
☐ B. Internal Validity
☐ C. Random Sample

□ D. Targetability

```
#Answer: A. External Validity.
```

8e. [3 points]Justify or refute the researcher's claims in 7d. Could the research done in this study apply to the population in San Antonio? List 3 reasons, in context, why or why not this may be the case.

8f. [1 point][bonus]When collecting these data, researchers made the following disclaimer.

These data were collected with respect towards personal customs, creeds, and with knowing consent from all participating subjects. All participants had the opportunity to view data collected on their behalf, and were offered a research-team led briefing on the results of the study. De-identified copies of these data and reports generated were made available at local universities.

Briefly explain two ethical principle of research that this disclaimer upholds and how actions taken by the research team applies to these ethical principles.

```
#Answer: 1) consent of participants -- obtained consent beforehand.
#Answer: 2) Data Sharing -- participants were able to view data collected
#Answer: 3) Beneficience -- study results were diseminated.
```

Question 9 is an optional question where we collect feedback on the exam.

9.[feedback] Take a moment to let us know any issues that came up on the exam.