PH142 Spring 2022 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.

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 students **here**

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

3.Please rename the file LASTNAME_FIRSTNAME_Midterm1_Spring2022.pdf

NOTES:

- 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*

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

Problem 1: 9 points

Problem 2: 5 points + 1 bonus point

Problem 3: 7 points

Problem 4: 8 points

Total: 30 points + 1 bonus point

Question 1 [9 points total]

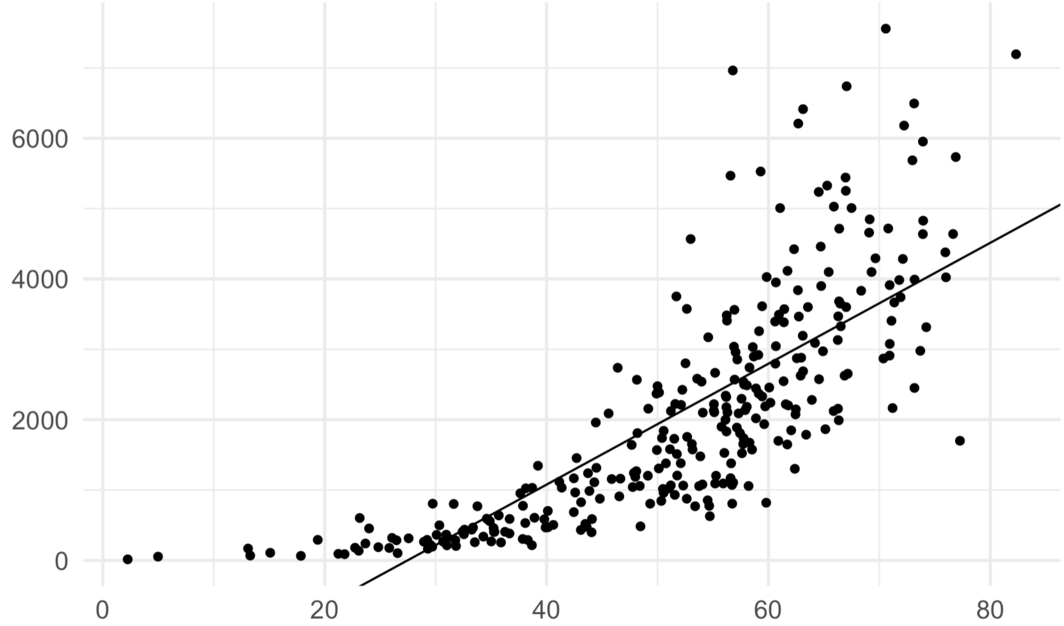
Researchers from a health insurance company have hired you to help analyze data about the relationship between the age of their enrollees (in years) and Medicare claim expenses (in USD $). An excerpt of the data is shown below.

```
## # A tibble: 6 x 3
##     AGE CLAIM_AMT  DIAB
##   <dbl>     <dbl> <dbl>
## 1    19       300     0
## 2    25       450     1
## 3    39       500     1
## 4    45       470     0
## 5    22       260     1
## 6    65      3000     0
```

1.1 [2 points] Which variable is the explanatory variable and which is the response variable? Explain your reasoning in 1-2 sentences.

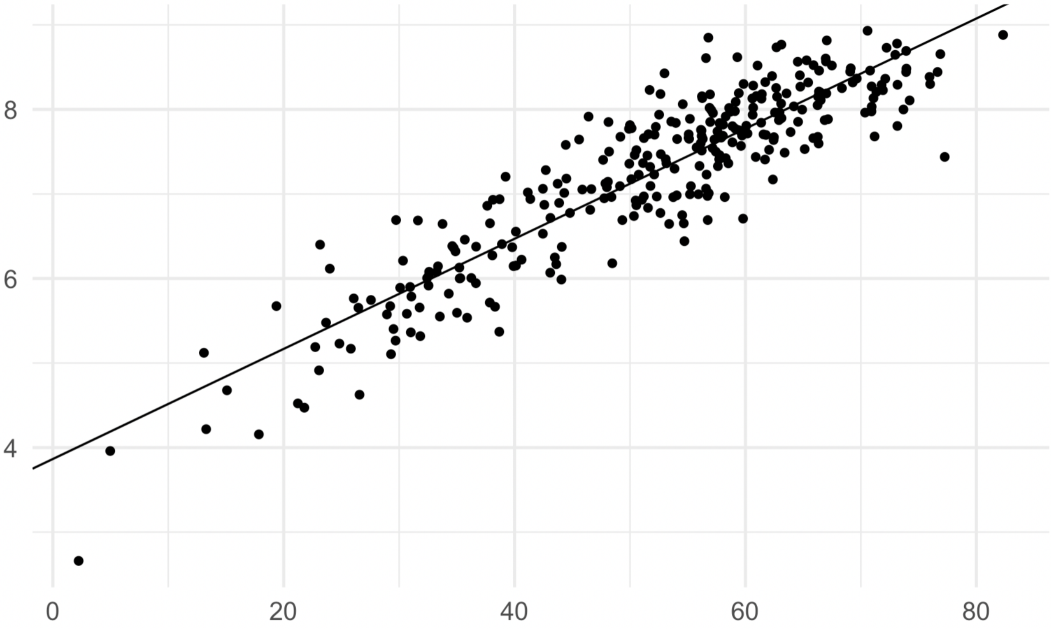
SOLUTION: explanatory is age, response is the claim amount in USD. It makes sense that as age increases, people would have to pay more insurance money (maybe they get sicker and go to the doctor more often). It doesn't make sense that insurance claims money would be the explanatory variable and that age would result from this, since people age regardless of insurance claims paid.

1.2 [1 point] Next, you use a scatter plot to visualize these data. Describe the relationship between age and Medicare claim expenses based on the plot below.



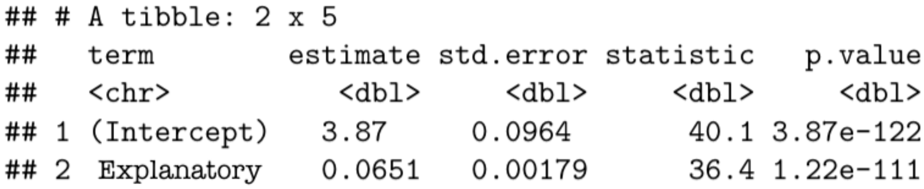
SOLUTION: moderately strong positive linear relationship. No outliers. Also okay to say non-linear.

1.3 [2 points] You perform a transformation on your data to test whether it will improve the strength of the relationship. Which variable(s) did you transform? How do you know?



SOLUTION: natural log transform medicare expenses (outcome). If you pick a point on the original plot, such as (60, 3000), you can find where that point would be on the transformed plot. You would still look at the same age (60) and find that the point on the line would be around 8. log(3000) = 8 so we know we took the natural log of the medicare expenses variable.

1.4 [1 point] You run a linear model on these data and are shown the output below. Interpret the slope in the context of this question.



```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   3.87     0.0964       40.1 3.87e-122
## 2  Explanatory  0.0651   0.00179      36.4 1.22e-111
```

SOLUTION: for every one year increase in age, there is an average of a 0.065 increase in the natural log of US dollars spent on medicare claims for enrollees in this insurance company.

1.5 [1 point] Interpret the intercept in the context of this question.

SOLUTION: One who is 0 years of age would have an insurance claim charge of the natural log of 3.87 USD.

1.6 [2 points] The insurance company asks you to predict the medicare claim expense for someone who is 95 years old. Calculate the predicted charge and explain whether you think this is a reasonable prediction. Only round your final answer (not the interim steps) to two decimal places.

Predicted charges:

Explanation:

SOLUTION: 276.85 –> 277 calculation: log(claim) = 0.06513(age) + 3.865 log(claim) = 0.06513*95 + 3.865 = 10.05 exp(log(claim)) = exp(10.05) claim = 23210.27

This is extrapolation - may not be reasonable since we don't have data for people who are around 95 years old

Question 2 [5 points total + 1 bonus point]

You are part of a team investigating the mental health of elderly, rural Japanese people. Your PI is interested in studying whether counseling will improve the mental health of this population, so she proposes an intervention where towns are assigned to an experimental or control group. Experimental groups would have clinics that offer weekly mental health group counseling where all individuals from that town would be eligible to participate. Control groups would offer a walking club in lieu of the mental health counseling.

2.1 [2 points] What kind of problem is your PI interested in addressing? Use the PPDAC framework and explain your choice.

SOLUTION: this is a causative/etiologic question. The PI wants to implement an experimental study to determine if counseling can effectively improve the mental health of the population. We are not predicting the mental health of the population nor are we just describing characteristics of the population, so it is neither a predictive nor a descriptive study.

2.2 [1 point] What unit of randomization is your PI proposing for this study?

SOLUTION: towns. The study design is such that towns are selected to have an intervention/not have the intervention. Therefore, the unit of randomization is at the township level.

2.3 [2 points] As part of the experimental design, your team develops a data sharing agreement. Local activists have insisted that this agreement include a clause that states a) that the data ultimately belongs to the study participants, b) participants can drop out of the study without consequence, and c) any research findings be translated and disseminated to the public. Explain an ethical principle that these activists are addressing in 1-2 sentences.

SOLUTION: Some combination of participant ownership of data and distribution of information to the public for the public's/individual's benefit (i.e., the principle of Justice) or beneficence (i.e., benefits of research should go to participants). Data belonging to the participants means they can drop out any time and not worry that their data being compromised.

Other solution may be that informed consent can be revoked at any time. Participants can drop out of study without consequence.

BONUS [1 point] Explain one reason why your PI chose her proposed unit of randomization.

SOLUTION: If randomization was at the individual level, there would be more potential for contamination - in other words, perhaps people in the counseling group would tell their friends who were randomized to the control group about the counseling. Then some people who were supposed to be in the walking group would then receive treatment and the inference from our results wouldn't be as strong.

Question 3 [7 points total]

The dean of the School of Public Health at a recently-established university in the Bay Area has recruited some students from UC Berkeley (you!) to perform exploratory data analysis on their recent recruitment statistics (why the dean didn't ask students from his own school, who knows). The dean is especially interested in the number of acceptances across different regions of the United States and Internationally. After cleaning and consolidating all of the data on the 5341 total admissions, you are left with the dataframe below.

```
##            region         department admits
## 1  International             Epibio    245
## 2        West US             Epibio    241
## 3     Midwest US             Epibio    235
## 4       South US             Epibio    233
## 5    Northeast US            Epibio    240
## 6  International   Community Health    230
## 7        West US   Community Health    226
## 8     Midwest US   Community Health    219
## 9       South US   Community Health    216
## 10  Northeast US   Community Health    222
## 11 International      Health Policy    145
## 12       West US      Health Policy    142
## 13    Midwest US      Health Policy    136
## 14      South US      Health Policy    132
## 15  Northeast US      Health Policy    130
## 16 International Infectious Diseases    465
## 17       West US Infectious Diseases    491
## 18    Midwest US Infectious Diseases    461
## 19      South US Infectious Diseases    460
## 20  Northeast US Infectious Diseases    472
```

A data dictionary is provided below for your convenience:

Column	Description
region	Region of the US (or international) applicants were from
department	Department that the applicants were accepted to
admits	Number of applicants who were accepted in a particular region and department

3.1 [1 point] What percentage of the acceptances are from the Western United States? Round to two decimal places.

SOLUTION: 1100 / 5341 = 0.20595 = 20.60%

3.2 [2 points] What is the marginal distribution of department acceptances? Round to two decimal places.

Epi/Bio: □

Community Health: □

Health Policy: □

Infectious Disease: □

SOLUTION: Epi/Bio: 1194/5341 = 0.2235536 = 22.36% Community Health: 1113/5341 = 0.2083879 = 20.84% Health Policy: 685/5341 = 0.1282531 = 12.82% Infectious Disease: 2349/5341 = 0.4398053 = 43.98%

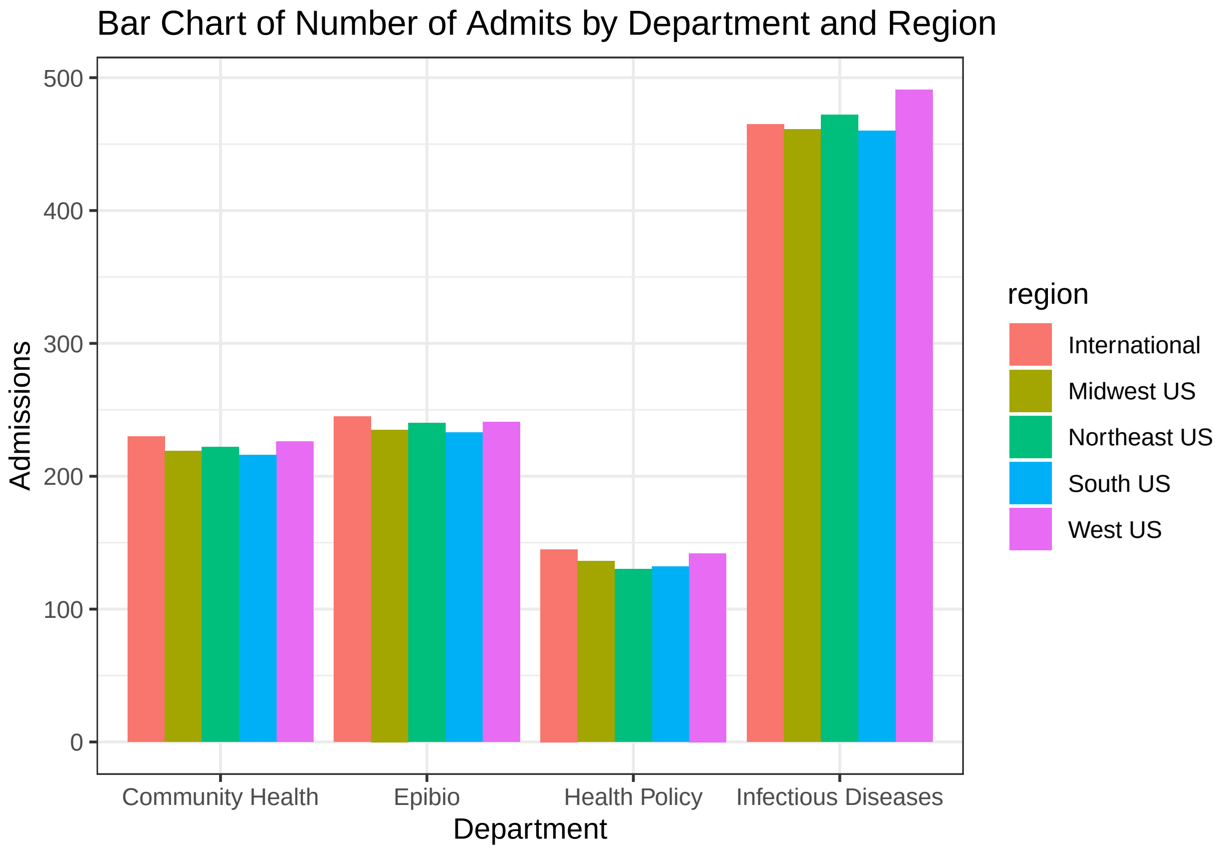
3.3 [1 point] Of the acceptances to the Health Policy program, what percentage are from the Northeastern US? Round to two decimal places.

□

SOLUTION: 130 / 685 = 0.1898 = 18.98%

Your peer decides to create a visualization to show how many admits each department has for each region. Her plot is shown below.

```
ggplot(sph_admits, aes(x = department, y = admits)) +
  geom_bar(aes(fill = region), stat="identity", position="dodge") +
  labs(title = "Bar Chart of Number of Admits by Department and Region", x = "Department", y = "Admissi
  theme_bw()
```



3.4 [1 point] Fill in the blanks to the ggplot code to recreate the plot above. Be sure to use the correct variable names from the dataframe above.

```
ggplot(sph_admits, aes(x = ____A____, y = ____B____)) +
  geom_bar(aes(fill = ____C____), stat = ____D____, position="____E____") +
  labs(title = "Number of Admits by Department, Dodged by Region", x = "Department", y = "Admissions") +
  theme_bw()
```

A: □

B: □

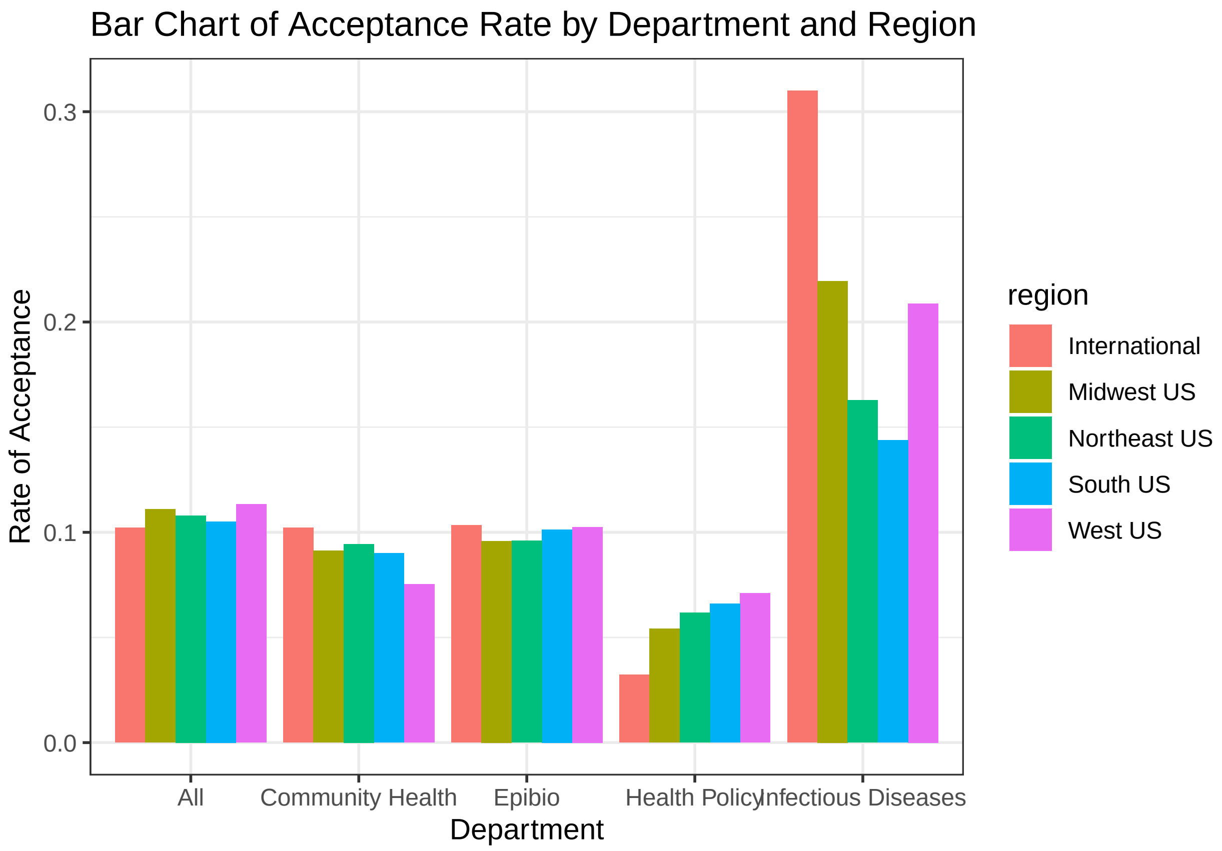
C: □

D: □

E: □

SOLUTION: A: department B: admits C: region D: “identity” E: dodge

From a quick glance at the bar graph above, it seems like departments across the School of Public Health are accepting roughly similar numbers of students from each region category. Seems like great news! However, you realized that the dean only included the number of acceptances, not the statistics on the number of applicants from each region. You did a bit of digging, and found some data on the number of students from each region who applied to each program. You use this new data to calculate the admissions rate and plot your results below.



3.5 [2 points] **From the graph above, it seems like international applicants have the highest acceptance rate for most departments compared to the other regions of the US. However, international applicant acceptance rates are the lowest overall. Name this phenomenon and explain how it might occur in this scenario.**

SOLUTION: Simpson's Paradox - We can see that the acceptance rates for International applicants are higher in every department than any other regions (except for Health Policy), and yet international applicants have the lowest acceptance rates overall. One explanation for this could be more international applicants applying to departments with the least acceptance rates (i.e. Health Policy) and fewer international applicants applying to the departments with higher acceptance rates (e.g. Infectious Diseases) basically the number of applicants is the confounding variable.

Question 4 [8 points total]

PrEP (Pre-exposure Prophylaxis) is a drug that helps prevent the transmission of HIV for at-risk populations. Researchers are interested in investigating the effect of time spent on PrEP (in months) on the onset of kidney failure. They collect a sample of 400 individuals, of whom 200 are classified as having kidney failure and 200 are not. Researchers collected information from consenting individuals on PrEP and put the following variables in a dataset called `prep_data`.

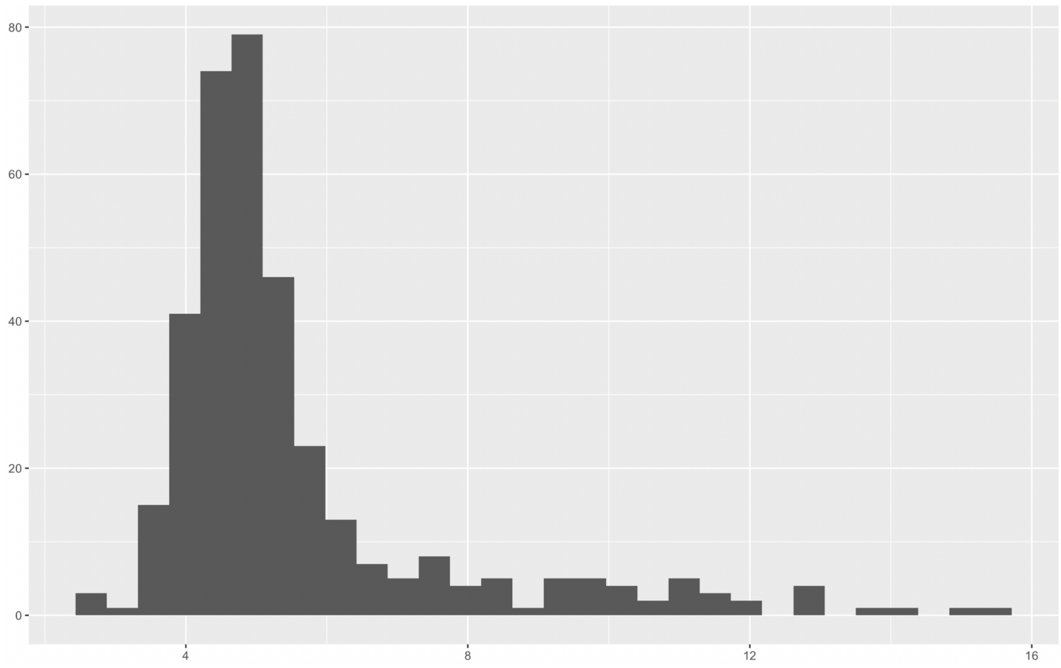
`prep_time`: time (months) on prep

`gfr`: glomerular filtration rate (millilters/min), a measure of kidney function

`age`: participant's age (years)

`id`: participant's identification number

4.1 [1 point] Describe the distribution of the `prep_time` variable using the histogram below.



SOLUTION: shape = skewed right center: mean around 6-7 months or median around 5-6 months spread: from 1 month to 15 months no distinct outliers

4.2 [1 point] The investigators then run a linear regression model of the `gfr` variable regressed on the `prep_time` variable and obtain the results below. Interpret the R^2 value in the context of this problem.

```
## # 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     0.816         0.816 0.452     1324. 1.22e-111     1  -186.  379.  390.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

SOLUTION: 81.6% of the variation in the glomerular filtration rate (kidney failure) is explained by the variation in PrEP usage in months.

4.3 [2 points] The researchers hypothesize that another variable may be influencing the relationship between kidney failure rates and time spent on PrEP. Give an example of a variable that may be influencing the relationship and justify your reasoning.

SOLUTION: Age is a potential confounding variable. Age may influence, explain both exposure and outcome.

4.4 [2 points] The team decides that they would like to subset their sample to include only individuals over 50 years of age and calculate the mean `gfr_ratio`, a variable that represents the ratio of glomerular filtration rate to the months spent on PrEP. Write the line of code that subsets your original sample to the individuals of interest, creates the `gfr_ratio` variable, calculates the mean of this variable, called `mean_gfr_ratio`, and assign this to a dataframe called `mean_prep_over_50`.

SOLUTION: mean_prep_over_50 <- prep_data %>% filter(age > 50) %>% mutate(gfr_ratio = gfr/prep_time) %>% summarize(mean_gfr_ratio = mean(gfr_ratio))

4.5 [2 points] Can researchers generalize the sample distribution of kidney failure to the general population? Why or why not?

SOLUTION: No, the researchers cannot generalize the marginal distribution of kidney failure because they collected their sample conditional on kidney failure as the outcome.

Exam feedback:

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

END OF EXAM