Homework 10

Your name and student ID

Today's date

- Solutions will be released on Nov 24.

Section 1: Voting during the 1992 election [21 points]

In the spirit of the upcoming 2020 presidential election, I thought it would be interesting to consider some historical data on voting patterns across US counties.

This code loads in the data frame `counties`:

```r
load("A10_counties.sav")
```

These data are from the 1992 election and looks at the percent of votes cast (for each county) for the `democrat` (Bill Clinton), `republican` (George Bush), and independent presidential nominees (Ross `Perot`).

Ideally, if you were interested in voting patterns, you might look at the relationship between individual characteristics and whether each individual voted Democrat or Republican. However, data like that is often hard to come by. The `counties` data provide data on 3141 counties. Use `View()` to examine these data briefly and read the labels corresponding to the variables. Note that Alaska is not included and that two other counties with populations = 0 have also been excluded.

As discussed in class we have the entire population (not just a sample), so strictly speaking we don't need to perform statistical inference. However, we might pretend this is a sample so that we can apply the techniques of inference and gain competence creating and interpeting a linear model.

1. [2 points] Looking only at California, plot the relationship between the % of votes cast for the Democratic candidate (`democrat`) and the population density of the county (`pop.density`). Since we will only be using the counties from California, go ahead and subset the full `counties` dataset to only include observations from the state of CA.

```r
counties_CA <- "SUBSET DATA HERE"

p1 <- "YOUR PLOT HERE"
p1
```

```
## [1] "YOUR PLOT HERE"
```

```r
check_problem1()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot for p1."
## [1] "Checkpoint 2 Error: Did you plot the right variable on the x axis for p1?"
## [1] "Checkpoint 3 Error: Did you plot the right variable on the y axis for p1?"
## [1] "Checkpoint 4 Error: Did you define a scatterplot in p1?"
##
## Problem 1
## Checkpoints Passed: 0
## Checkpoints Errored: 4
## 0% passed
## --------
## Test: FAILED
```

2. [1 point] The above plot you made does not look very good. The distribution of population density is skewed right, with a few counties having much higher densities than the majority of counties. To see which counties these are, we will use `geom_text_repel` from the library `ggrepel` (loaded at the top of this assignment). The template for using this function is: `geom_text_repel(aes(label = your_labelling_var))`. You will want to set the labeling variable to be the variable in the dataset containing the county names.

```
p2 <- "YOUR CODE HERE"
p2
```

```
## [1] "YOUR CODE HERE"
```

```
check_problem2()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot for p2."
## [1] "Checkpoint 2 Error: Did you plot the right variable on the x axis for p2?"
## [1] "Checkpoint 3 Error: Did you plot the right variable on the y axis for p2?"
## [1] "Checkpoint 4 Error: Did you define a scatterplot in p14?"
## [1] "Checkpoint 5 Error: Did you add labeling to p2?"
##
## Problem 2
## Checkpoints Passed: 0
## Checkpoints Errored: 5
## 0% passed
## --------
## Test: FAILED
```

The current issue with these data is that San Francisco (as you can now hopefully point out) has a much higher population density than other counties, and that generally there is a large right skew in the distribution of the population density variable.

If we tried and fit a linear model to these data, it would not fit well– because the relationship between population density and the response variable is not linear. However, this is the perfect situation to try transforming the x variable.

3. [2 points] Try adding a log-transformed version of population density to the data frame and remake your plot using this new variable. Call this new variable `log_pop_density`. Keep the population labels. Also add a smoothed fitted line:

```
# uncomment the line below by deleting the pound sign
# counties_CA <- "Add new variable here"

p3 <- "YOUR PLOT HERE"
p3
```

```
## [1] "YOUR PLOT HERE"
```

```
check_problem3()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot for p3."
## [1] "Checkpoint 2 Error: Did you plot the right (and correctly named) variable on the x axis for p3?
## [1] "Checkpoint 3 Error: Did you plot the right variable on the y axis for p3?"
## [1] "Checkpoint 4 Error: Did you define a scatterplot in p3?"
## [1] "Checkpoint 5 Error: Did you add labeling to p3?"
##
## Problem 3
## Checkpoints Passed: 0
## Checkpoints Errored: 5
## 0% passed
## --------
## Test: FAILED
```

4. [4 points] Describe the relationship between the (logged) population density and the response variable in terms of the shape, direction, strength, and outliers. These are concepts from Chapter 3. Calculate the correlation (round to 4 decimals) to comment on one of these aspects.

```
p4 <- "CALCULATE THE CORRELATION HERE"
p4
```

```
## [1] "CALCULATE THE CORRELATION HERE"
```

```
check_problem4()
```

```
## [1] "Checkpoint 1 Error: Incorrect"
## 
## Problem 4
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

[TODO: YOUR ANSWER HERE]

5. [4 points] Run a linear model regression of the % votes cast for the democratic candidate as a function of the population density. Make sure you get the order of variables right in the `lm()` function! Use the `tidy()` function to show the slope and intercept estimates. Interpret the relationship between the logged population density and the response variable. (You can `View()` the data frame to make sure you are getting your units right by checking the descriptions in the labels for each variable). Use another function from `broom` show the r-squared. Report and interpret this value for the model.

```r
lm_CA <- "YOUR MODEL HERE"

r.squared <- "Report r-squared here. Leave as decimal and round to 2 places"


check_problem5()
```

```
## [1] "Checkpoint 1 Error: Did you make a linear model for lm_CA"
## [1] "Checkpoint 2 Error: Did you calculate r-squared?"
##
## Problem 5
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

[TODO: YOUR ANSWER HERE]

6. [4 points] Using the code learned in class, that was also shown in Lab 11, make the four plots to examine the assumptions.

```r
plot1 <- "Code for scatterplot here"
plot1
```

```
## [1] "Code for scatterplot here"
```

```r
plot2 <- "Code for QQ plot here"
plot2
```

```
## [1] "Code for QQ plot here"
```

```r
plot3 <- "Code for Fitted vs. Residuals plot here"
plot3
```

```
## [1] "Code for Fitted vs. Residuals plot here"
```

```r
plot4 <- "Code for Amount explained plot here"
plot4
```

```
## [1] "Code for Amount explained plot here"
```

```r
check_problem6()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot for plot1."
## [1] "Checkpoint 2 Error: Did you plot the right variable on the x axis for plo1?"
## [1] "Checkpoint 3 Error: Did you plot the right variable on the y axis for plot1?"
## [1] "Checkpoint 4 Error: Did you define a scatterplot in plot1?"
## [1] "Checkpoint 5 Error: You did not define a ggplot for plot2."
## [1] "Checkpoint 6 Error: Did you plot the right variable in ’sample’ for plot2"
## [1] "Checkpoint 7 Error: Did you define a QQplot in plot2?"
## [1] "Checkpoint 8 Error: You did not define a ggplot for plot3."
## [1] "Checkpoint 9 Error: Did you plot the right variable on the x axis for plot3?"
## [1] "Checkpoint 10 Error: Did you plot the right variable on the y axis for plot3?"
## [1] "Checkpoint 11 Error: Did you define a scatterplot in plot3?"
## [1] "Checkpoint 12 Error: You did not define a ggplot for plot4."
## [1] "Checkpoint 13 Error: Did you define a boxplot in plot4?"
##
## Problem 6
## Checkpoints Passed: 0
## Checkpoints Errored: 13
## 0% passed
## --------
## Test: FAILED
```

7. [4 points] Comment on each of the plots and conclude about which assumptions appear violated vs. not violated. Don’t forget to comment on the one assumption that cannot be investigated using plots.

[TODO: YOUR ANSWER HERE]

Section 2: Abstract interpretation [5 points]

Read the following abstract and answer the questions that follow. J Asthma. 2018 Oct 11:1-12. doi: 10.1080/02770903.2018.1508471. [Epub ahead of print] Impact of scenario based training on asthma first aid knowledge and skills in school staff: an open label, three-arm, parallel-group repeated measures study. Luckie K1, Saini B1, Soo YYB1,2, Kritikos V1,3, Charles Collins JB1, Jane Moles R1.

OBJECTIVE: To test the hypothesis that scenario-based skills training is more effective than knowledge training alone in improving the asthma first aid (AFA) skills of school personnel. Education developed specifically for non-primary caregivers such as school staff is vital to minimize the risk of mortality associated with asthma.

METHODS: Schools were allocated to one of three arms to compare AFA knowledge and AFA skills. Arm 1 underwent conventional asthma training, arm 2 underwent scenario-based training and arm 3 had a combination of the two. Conventional asthma training involved a didactic oral presentation. The scenario-based skills training required the participant to describe and demonstrate how they would manage a child having a severe exacerbation of asthma using equipment provided. Follow-up occurred at 3 weeks post baseline and again between 3-7 months after the first training/education visit.

RESULTS: Nineteen primary schools (204 participants) were recruited. One-way ANOVA and Bonferroni Post-Hoc Tests showed there was a significant difference in AFA skills scores between the study arms who underwent scenario-based training; arms 2 and 3 (91.5% and 91.1%) and arm 1 who underwent conventional asthma training (77.3%) ($p<0.001$). AFA knowledge improved significantly in all study arms with no differences between study arms. Improvements seen in both AFA knowledge and AFA skills were maintained over time.

CONCLUSIONS: Scenario-based training was superior to conventional didactic asthma training for AFA skills acquisition and overall competency in the administration of AFA and should be included in future asthma training programs.

8. [1 point] Two methods of hypothesis testing (types of tests) are mentioned in the abstract. What is the null hypothesis for each of these tests (please list in the order they are mentioned in the abstract?

H_0: [TODO: YOUR ANSWER HERE] - one sentence only

H_0: [TODO: YOUR ANSWER HERE] - one sentence only

9. [1 point] There are two outcomes of interest in this study. For which **outcome** would you conclude that there is a significant difference between the training groups.

[TODO: YOUR ANSWER HERE]

10. [1 point] If you were a school administrator why might you choose the arm 3 training?

[TODO: YOUR ANSWER HERE]

11. [1 point] List one question you might want to ask about the methods, sample or results that would help you interpret the findings of this study?

[TODO: YOUR ANSWER HERE]

12. [1 point] What is another test that could have been considered for these study data?

[TODO: YOUR ANSWER HERE]

Section 3: ANOVA and Tukey's HSD [6 points]

Note: This material will be taught on Monday, November 23rd

For this question we will use the data from the NHANES survey ‘

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   ridageyr = col_double(),
##   drinks = col_double(),
##   bmxwt = col_double(),
##   bmxht = col_double(),
##   bmxbmi = col_double(),
##   bpxpls = col_double(),
##   bpxsy1 = col_double(),
##   bpxsy2 = col_double(),
##   bpxdi1 = col_double(),
##   bpxdi2 = col_double(),
##   lbdhdd = col_double(),
##   sleep = col_double(),
##   lbdldl = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: ‘...‘ is not empty.
##
## We detected these problematic arguments:
## * ‘needs_dots‘
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 6 x 40
##   ridageyr agegroup gender military born  citizen drinks drinkscat bmxwt bmxht
##      <dbl> <chr>    <chr>  <chr>    <chr> <chr>    <dbl> <chr>     <dbl> <dbl>
## 1       72 65+      Male   History~ Born~ US cit~      0 0          88.9  175.
## 2       73 65+      Female No       Born~ US cit~      0 0          52    162.
## 3       61 50-64    Female No       Born~ US cit~      2 11-Jan     93.4  162.
## 4       26 20-34    Female No       Born~ US cit~    209 96-364     47.1  152.
## 5       33 20-34    Female No       No    US cit~     NA <NA>       56.8  158
## 6       32 20-34    Male   No       No    No         300 96-364     79.7  166.
## # ... with 30 more variables: bmxbmi <dbl>, bmicat <chr>, bpxpls <dbl>,
## #   bpxsy1 <dbl>, bpxsy2 <dbl>, sys1d <chr>, sys2d <chr>, bpxdi1 <dbl>,
## #   bpxdi2 <dbl>, dias1d <chr>, dias2d <chr>, bpcat <chr>, chest <chr>,
## #   fs1 <chr>, fs2 <chr>, fs3 <chr>, lbdhdd <dbl>, hdlcat <chr>, highhdl <chr>,
## #   hi <chr>, asthma <chr>, vwa <chr>, vra <chr>, va <chr>, aspirin <chr>,
## #   sleep <dbl>, is <chr>, hs <chr>, lbdldl <dbl>, highldl <chr>
```

13. [1 point] Generate the mean and standard deviations in a dataframe for blood lipid level “lbdldl” by Blood pressure group “bpcat”. Use dplyr functions.

```
p13 <- "Your code here"
p13
```

```
## [1] "Your code here"
```

```
check_problem13()
```

```
## [1] "Checkpoint 1 Error: p13 should be a dataframe"
## [1] "Checkpoint 2 Error: Wrong answer for mean"
## [1] "Checkpoint 3 Error: wrong answer for sd"
##
## Problem 13
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

14. [1 point] Create a boxplot that helps you to visualize these data.

```
p14 <- "Your plot here"
p14
```

```
## [1] "Your plot here"
```

```
check_problem14()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot."
## [1] "Checkpoint 2 Error: Did you use the right dataset?"
## [1] "Checkpoint 3 Error: Did you plot the right variable on the x axis?"
## [1] "Checkpoint 4 Error: Did you plot the right variable on the y axis?"
## [1] "Checkpoint 5 Error: Did you define a boxplot in ggplot?"
##
## Problem 14
## Checkpoints Passed: 0
## Checkpoints Errored: 5
## 0% passed
## --------
## Test: FAILED
```

15. [2 points] Conduct an ANOVA with Tukey's HSD for these data. Assign your model to the variable `tukey`.

```
tukey <- "Your code here"
p15 <- tidy(tukey) #keep this line


check_problem15()
```

```
## [1] "Checkpoint 1 Passed: Correct!"
## [1] "Checkpoint 2 Error: Did you use the correct formula"
## [1] "Checkpoint 3 Error: Did you use the correct formula"
## [1] "Checkpoint 4 Error: Incorrect estimates. Check your code"
##
## Problem 15
## Checkpoints Passed: 1
## Checkpoints Errored: 3
## 25% passed
## --------
## Test: FAILED
```

16. [1 point] What are the null and alternative hypotheses for this test?

[TODO: YOUR ANSWER HERE]

17. [1 point] What do you conclude from your analysis?

[TODO: YOUR ANSWER HERE]

Section 3: Non-parametric [3 points]

Note: This material will be taught on Monday, November 30th

You are testing the change in test scores following an intensive tutoring session.
You have the following data from a small group of students each student is tested before and after the tutoring session.
Each row represents one student.

Time 1	Time 2
65	77
87	100
77	75
90	89
70	80
84	81
92	91
83	96
85	84
91	89
68	88
72	100
81	81

```r
#this code makes a dataframe of the table you see above
test_scores <- tribble(
  ~time1, ~time2,
  65, 77,
  87, 100,
  77, 75,
  90, 89,
  70, 80,
  84, 81,
  92, 91,
  83, 96,
  85, 84,
  91, 89,
  68, 88,
  72, 100,
  81, 81)
```

18. [2 point] Calculate the appropriate non-paramentric test for these data by hand. Attach an image to show your work. Make sure to place the image in the `src` directory. Uncomment the line by deleting the pound sign. Report the p-value by saving it p18. Keep it as a decimal and round to 4 places.

```r
#knitr::include_graphics("src/path-to-file")
p18 <- "YOUR P-VALUE HERE"
p18
```

```
## [1] "YOUR P-VALUE HERE"
```

```r
check_problem18()
```

```
## [1] "Checkpoint 1 Error: Incorrect p-value. Check your work"
##
## Problem 18
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

19. [1 point] Check your work using [insert your test].test() function in R. Keep your answer as a decimal rounded to 4 decimals. Report your p-value and save it to the variable p19.

```r
p19 <- "Your p-value here"
p19
```

```
## [1] "Your p-value here"
```

```r
check_problem19()
```

```
## [1] "Checkpoint 1 Error: Incorrect p-value. Check your work"
##
## Problem 19
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

Check your score

Click on the middle icon on the top right of this code chunk (with the downwards gray arrow and green bar) to run all your code in order. Then, run this chunk to check your score.

```
# Just run this chunk.
total_score()
```

```
##                       Test Points_Possible          Type
## Problem 1           FAILED               2    autograded
## Problem 2           FAILED               1    autograded
## Problem 3           FAILED               2    autograded
## Problem 4           FAILED               1    autograded
## Problem 5           FAILED               4    autograded
## Problem 6           FAILED               4    autograded
## Problem 7  NOT YET GRADED               4 free-response
## Problem 8  NOT YET GRADED               1 free-response
## Problem 9  NOT YET GRADED               1 free-response
## Problem 10 NOT YET GRADED               1 free-response
## Problem 11 NOT YET GRADED               1 free-response
## Problem 12 NOT YET GRADED               1 free-response
## Problem 13          FAILED               1    autograded
## Problem 14          FAILED               1    autograded
## Problem 15          FAILED               2    autograded
## Problem 16 NOT YET GRADED               1 free-response
## Problem 17 NOT YET GRADED               1 free-response
## Problem 18          FAILED               2    autograded
## Problem 19          FAILED               1    autograded
```