Homework 1B

STAT 242: Intermediate Statistics

Don't forget to load any packages using library() that you may need for this homework!

Problem 1: Adapted from *The Statistical Sleuth*, Problem 1.17

Seven students volunteered for a comparison of study guides for an advanced course in mathematics. They were randomly assigned to, four to study guide A and three to study guide B. All were instructed to study independently. Following a two-day study period, all students were given an examination about the material covered by the guides, with the following results:

- Study Guide A scores: 68 77, 82, 85
- Study Guide B scores: 53, 64, 71

(a) What is the difference between sample averages for the two groups? (1 pt)

(b) State relevant null and alternative hypotheses to compare the performance of the students assigned to each group. (2 pts)

You might use μ_1 and μ_2 or δ as your parameters. Whichever parameter(s) you choose to use in your statement of the hypotheses, please write a brief sentence defining what the parameter(s) are in the context of this example.

(c) There are 35 possible ways that these students could have been randomized to two groups, listed below. Using this information, perform a randomization test of the hypotheses you stated in part (b). For each such randomization, the difference between sample averages for the two groups is shown. Use this information to calculate the two-sided p value for the test. (2 pts)

```
Group A Scores: 68, 77, 82, 85       Group A mean: 78
Group B Scores: 53, 64, 71           Group B mean: 62.667
Group A Mean - Group B Mean: 15.333

Group A Scores: 68, 77, 82, 53       Group A mean: 70
Group B Scores: 85, 64, 71           Group B mean: 73.333
Group A Mean - Group B Mean: -3.333

Group A Scores: 68, 77, 82, 64       Group A mean: 72.75
Group B Scores: 85, 53, 71           Group B mean: 69.667
Group A Mean - Group B Mean: 3.083

Group A Scores: 68, 77, 82, 71       Group A mean: 74.5
Group B Scores: 85, 53, 64           Group B mean: 67.333
Group A Mean - Group B Mean: 7.167
```

```
Group A Scores: 68, 77, 85, 53       Group A mean: 70.75
Group B Scores: 82, 64, 71           Group B mean: 72.333
Group A Mean - Group B Mean: -1.583

Group A Scores: 68, 77, 85, 64       Group A mean: 73.5
Group B Scores: 82, 53, 71           Group B mean: 68.667
Group A Mean - Group B Mean: 4.833

Group A Scores: 68, 77, 85, 71       Group A mean: 75.25
Group B Scores: 82, 53, 64           Group B mean: 66.333
Group A Mean - Group B Mean: 8.917

Group A Scores: 68, 77, 53, 64       Group A mean: 65.5
Group B Scores: 82, 85, 71           Group B mean: 79.333
Group A Mean - Group B Mean: -13.833

Group A Scores: 68, 77, 53, 71       Group A mean: 67.25
Group B Scores: 82, 85, 64           Group B mean: 77
Group A Mean - Group B Mean: -9.75

Group A Scores: 68, 77, 64, 71       Group A mean: 70
Group B Scores: 82, 85, 53           Group B mean: 73.333
Group A Mean - Group B Mean: -3.333

Group A Scores: 68, 82, 85, 53       Group A mean: 72
Group B Scores: 77, 64, 71           Group B mean: 70.667
Group A Mean - Group B Mean: 1.333

Group A Scores: 68, 82, 85, 64       Group A mean: 74.75
Group B Scores: 77, 53, 71           Group B mean: 67
Group A Mean - Group B Mean: 7.75

Group A Scores: 68, 82, 85, 71       Group A mean: 76.5
Group B Scores: 77, 53, 64           Group B mean: 64.667
Group A Mean - Group B Mean: 11.833

Group A Scores: 68, 82, 53, 64       Group A mean: 66.75
Group B Scores: 77, 85, 71           Group B mean: 77.667
Group A Mean - Group B Mean: -10.917

Group A Scores: 68, 82, 53, 71       Group A mean: 68.5
Group B Scores: 77, 85, 64           Group B mean: 75.333
Group A Mean - Group B Mean: -6.833

Group A Scores: 68, 82, 64, 71       Group A mean: 71.25
Group B Scores: 77, 85, 53           Group B mean: 71.667
Group A Mean - Group B Mean: -0.417

Group A Scores: 68, 85, 53, 64       Group A mean: 67.5
Group B Scores: 77, 82, 71           Group B mean: 76.667
Group A Mean - Group B Mean: -9.167

Group A Scores: 68, 85, 53, 71       Group A mean: 69.25
```

```
Group B Scores: 77, 82, 64           Group B mean: 74.333
Group A Mean - Group B Mean: -5.083

Group A Scores: 68, 85, 64, 71       Group A mean: 72
Group B Scores: 77, 82, 53           Group B mean: 70.667
Group A Mean - Group B Mean: 1.333

Group A Scores: 68, 53, 64, 71       Group A mean: 64
Group B Scores: 77, 82, 85           Group B mean: 81.333
Group A Mean - Group B Mean: -17.333

Group A Scores: 77, 82, 85, 53       Group A mean: 74.25
Group B Scores: 68, 64, 71           Group B mean: 67.667
Group A Mean - Group B Mean: 6.583

Group A Scores: 77, 82, 85, 64       Group A mean: 77
Group B Scores: 68, 53, 71           Group B mean: 64
Group A Mean - Group B Mean: 13

Group A Scores: 77, 82, 85, 71       Group A mean: 78.75
Group B Scores: 68, 53, 64           Group B mean: 61.667
Group A Mean - Group B Mean: 17.083

Group A Scores: 77, 82, 53, 64       Group A mean: 69
Group B Scores: 68, 85, 71           Group B mean: 74.667
Group A Mean - Group B Mean: -5.667

Group A Scores: 77, 82, 53, 71       Group A mean: 70.75
Group B Scores: 68, 85, 64           Group B mean: 72.333
Group A Mean - Group B Mean: -1.583

Group A Scores: 77, 82, 64, 71       Group A mean: 73.5
Group B Scores: 68, 85, 53           Group B mean: 68.667
Group A Mean - Group B Mean: 4.833

Group A Scores: 77, 85, 53, 64       Group A mean: 69.75
Group B Scores: 68, 82, 71           Group B mean: 73.667
Group A Mean - Group B Mean: -3.917

Group A Scores: 77, 85, 53, 71       Group A mean: 71.5
Group B Scores: 68, 82, 64           Group B mean: 71.333
Group A Mean - Group B Mean: 0.167

Group A Scores: 77, 85, 64, 71       Group A mean: 74.25
Group B Scores: 68, 82, 53           Group B mean: 67.667
Group A Mean - Group B Mean: 6.583

Group A Scores: 77, 53, 64, 71       Group A mean: 66.25
Group B Scores: 68, 82, 85           Group B mean: 78.333
Group A Mean - Group B Mean: -12.083

Group A Scores: 82, 85, 53, 64       Group A mean: 71
Group B Scores: 68, 77, 71           Group B mean: 72
Group A Mean - Group B Mean: -1
```

```
Group A Scores: 82, 85, 53, 71       Group A mean: 72.75
Group B Scores: 68, 77, 64           Group B mean: 69.667
Group A Mean - Group B Mean: 3.083

Group A Scores: 82, 85, 64, 71       Group A mean: 75.5
Group B Scores: 68, 77, 53           Group B mean: 66
Group A Mean - Group B Mean: 9.5

Group A Scores: 82, 53, 64, 71       Group A mean: 67.5
Group B Scores: 68, 77, 85           Group B mean: 76.667
Group A Mean - Group B Mean: -9.167

Group A Scores: 85, 53, 64, 71       Group A mean: 68.25
Group B Scores: 68, 77, 82           Group B mean: 75.667
Group A Mean - Group B Mean: -7.417
```

(d) Write a sentence or two interpreting the p-value in context. (I am not looking for a conclusion for the test here, but a restatement of the definition of the p-value in the context of this example. We will talk about using p-values to draw conclusions in the next chapter.) (2 pts)

Problem 2: Adapted from Sleuth3 1.26

Each year, the League of Conservation Voters (LCV) identifies legislative votes taken in each house of the U.S. Congress – votes that are highly influential in establishing policy and action on evironmental problems. The LCV then publishes whether each member of Congress cast a pro-environment or an anti-environment vote. The following R code reads in a data set with the LCV's ratings for each member of House of Representatives from 2005 to 2007, and filters it to include results for only Democrats and Republicans (not Independents).

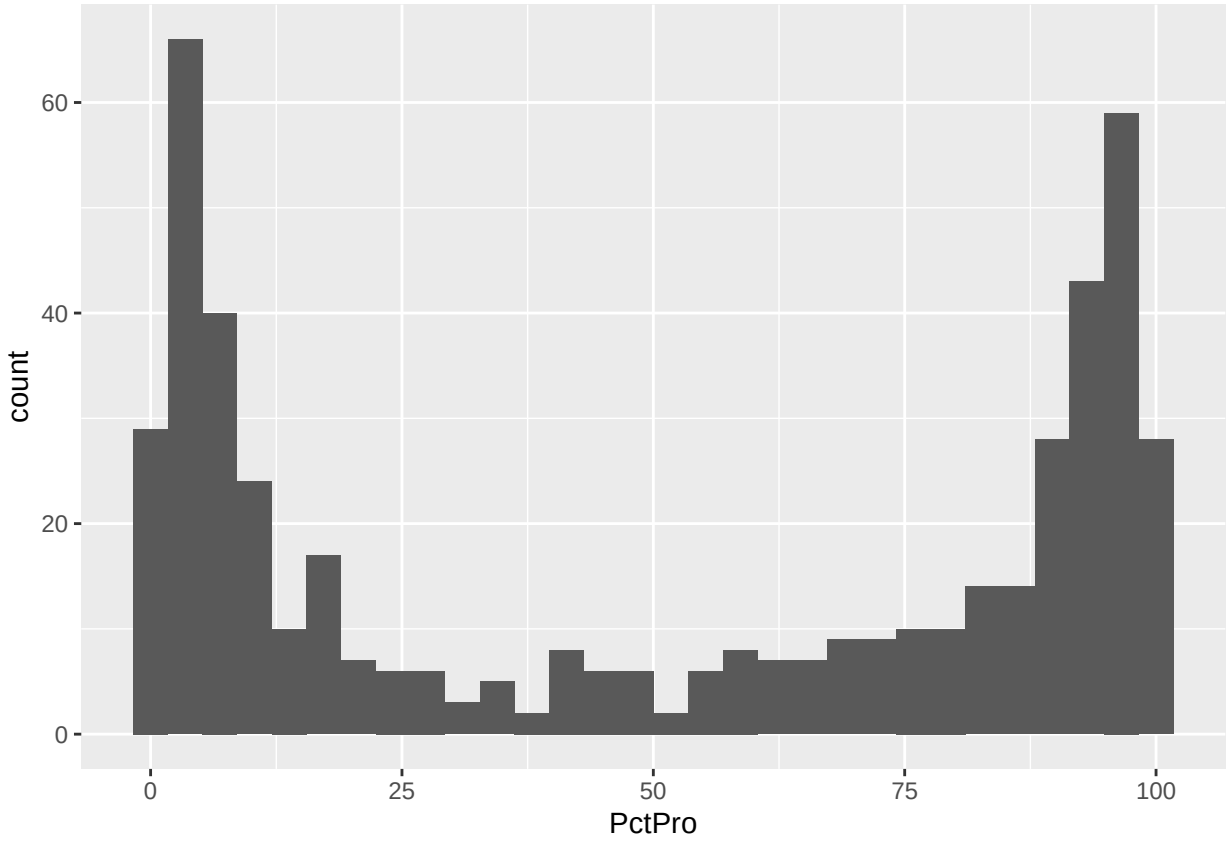
```
library(dplyr)
library(ggplot2)
library(Sleuth3)
lcv_ratings <- ex0126
lcv_ratings <- lcv_ratings %>%
  filter(
    Party %in% c("D", "R")
  )
```

(a) Variations on a histogram

The purpose of this part of the exercise is to explore a bunch of options for how to build histograms using `ggplot2` in R. As a starting point, here's a histogram summarizing the distribution of the overall percentage of "pro-environment" votes for all representatives in the data set:

```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Recall from previous labs/notes that the first line of this code sets up a new plot based on the `lcv_ratings` data frame. It also sets up an aesthetic mapping, putting the `PctPro` variable on the `x` (horizontal) axis. The geometry type of the plot is a `histogram`. Use this code as a starting point for the rest of part a.

i. Add facetting by the representatives' `Party` (0.5 pts) Add facetting to the plot code below using `facet_grid` or `facet_wrap` (either way, your choice).

ii. Add a fill by the representatives' `Party` (0.5 pts) Add a second aesthetic mapping by adding `, fill = Party` after `x = PctPro`.

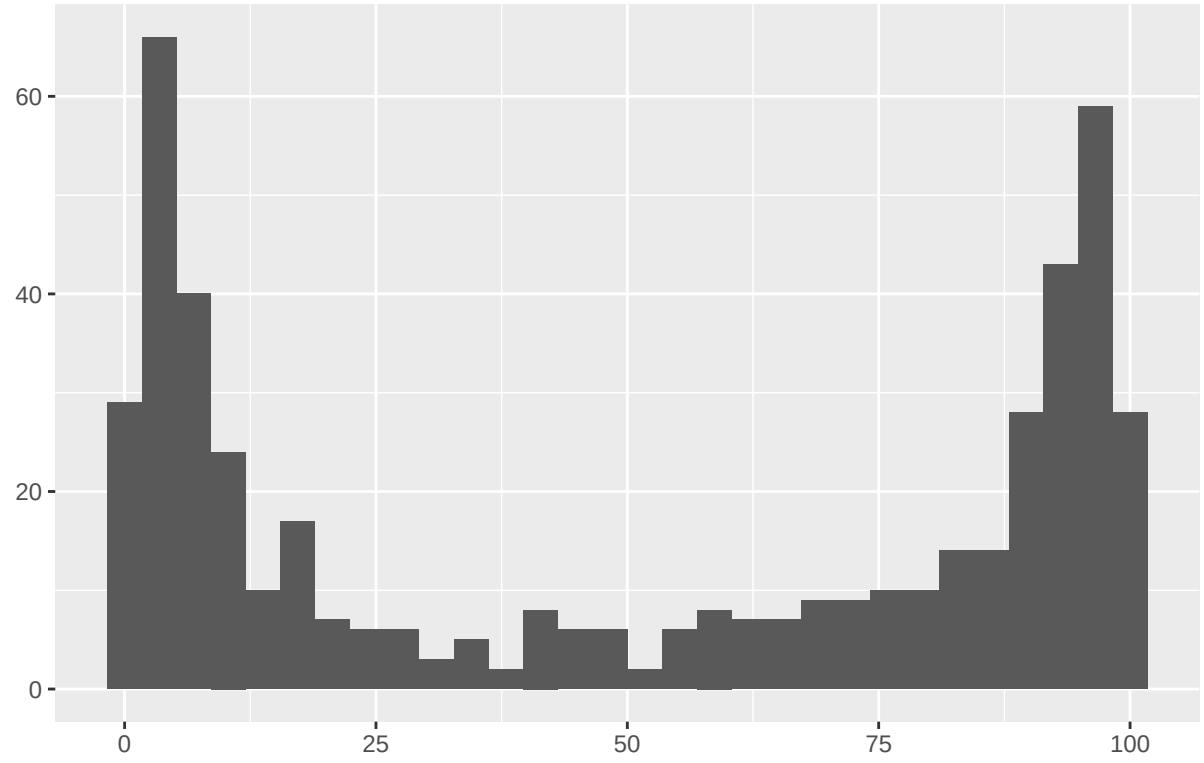
iii. Set the number of bins used and the boundary of the bins (0.5 pts) Add a `bins = 10` argument to the `geom_histogram()` function call (between the parentheses after `geom_histogram`). Experiment with the number of bins used until you find a number of bins that looks like it provides a good summary of the data.

Also add a `boundary = 0` argument to the `geom_histogram()` function call, separated by a comma from the `bins` argument. This argument specifies that one of the bins will have its left endpoint at 0 on the horizontal axis. The locations of the other bins are determined by the width of the bins.

iv. Set a more informative plot title and axis labels (0.5 pts) Using the ggtitle(), xlab(), and ylab() functions, you can change the graph title and axis labels. I've provided the set up for this below, but you should fill in appropriate values between the quotes.

```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro)) +
  geom_histogram() +
  ggtitle("") +
  xlab("") +
  ylab("")
```

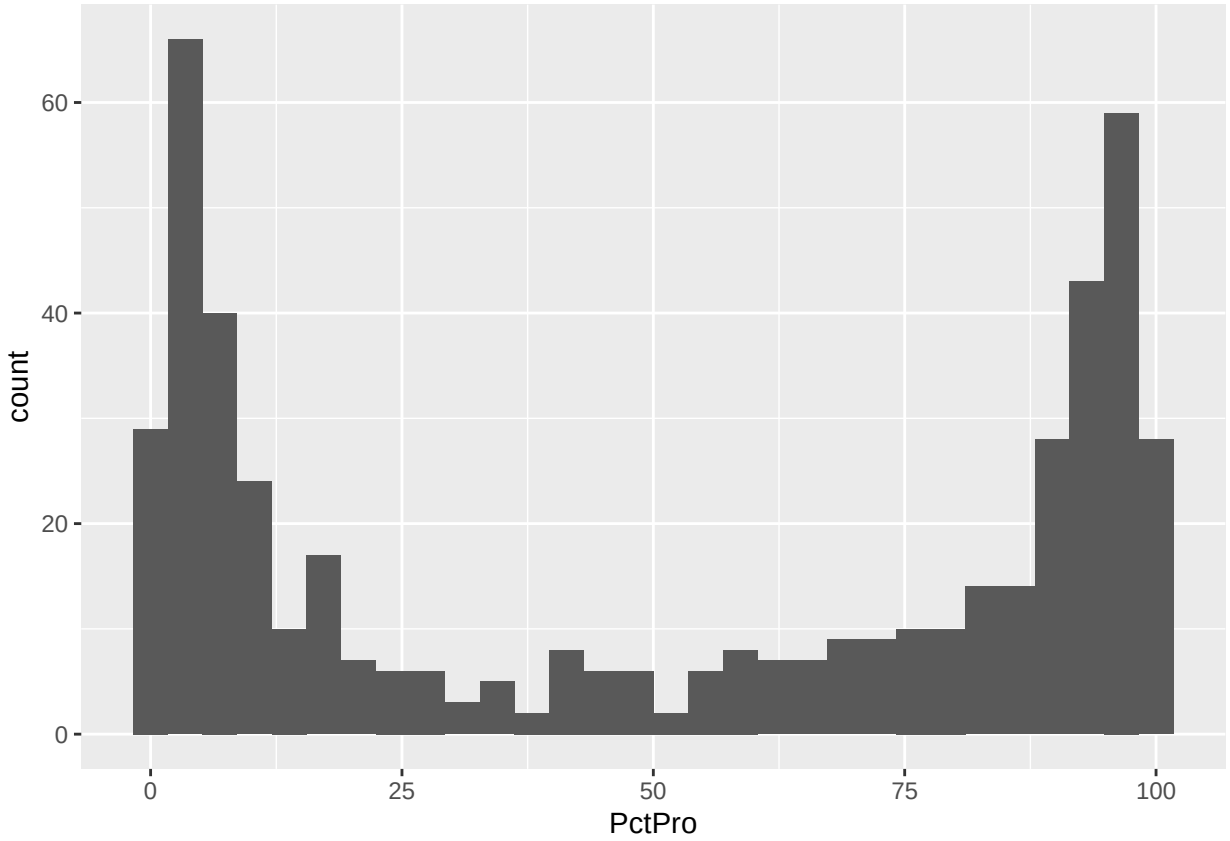
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = lcv_ratings, mapping = aes(x = PctPro)) +
  geom_histogram()
```

v. By combining your choice of some of the different ideas above, create a histogram that you think does the best job of conveying what's going on in the data set. (2 pts)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(b) Permutation test

Following subparts (i-iv) below, conduct a hypothesis test to compare pro-environment votes for Democratic and Republican members of the house of representatives.

i. Define the parameter or parameters you are making inference about. (1 pt)

ii. State the null and alternative hypotheses for a two-sided test. (2 pts)

```
library(mosaic)
# set a seed to get reproducible results
set.seed(513945)
# find observed difference in means for the actual sample data
group_means <- lcv_ratings %>%
  group_by(Party) %>%
  summarize(mean_PctPro = mean(PctPro))
observed_group_means_difference <- group_means$mean_PctPro[1] - group_means$mean_PctPro[2]
# simulate 10000 random assignments of legislators to the different parties,
# and for each simulation calculate the mean
simulation_results <- data.frame(
  group_means_difference = rep(NA, 1000)
```

```
)
for(i in seq_len(1000)) {
  shuffled_group_means <- lcv_ratings %>%
    mutate(
      Party = shuffle(Party)
    ) %>%
    group_by(Party) %>%
    summarize(mean_PctPro = mean(PctPro))

  simulation_results$group_means_difference[i] <- (shuffled_group_means$mean_PctPro[1] -
                                                     shuffled_group_means$mean_PctPro[2])
}
count_greater <- simulation_results %>%
  summarize(
    count = sum(group_means_difference >= observed_group_means_difference)
  )
count_less <- simulation_results %>%
  summarize(
    count = sum(group_means_difference <= -observed_group_means_difference)
  )
approximate_pval <- (count_greater$count + count_less$count) / 1000
approximate_pval
```

iii. The code below is adapted from the code in our first lab, and calculates an approximate p-value from a permutation test with 1,000 permutations. Run this code to get the approximate p-value (you don't need to modify the code at all). (Just so you're not confused by the output - the approximate p-value is 0!) (1 pt)

```
## [1] 0
```

iv. Write a sentence or two interpreting the p-value in context. (I am not looking for a conclusion for the test here, but a restatement of the definition of the p-value in the context of this example.) (3 pts)

(c) 2-sample *t*-test

Conduct a 2-sample *t*-test to compare pro-environment votes for Democratic and Republican members of the House of Representatives. Use the same parameter/parameters, null and alternative hypotheses that you used in the permutation test.

(i) Write R code to conduct the test, and report the test statistic and p-value from the statistical output. (2 pts)

(ii) Write a sentence or two interpreting the p-value in context. (2 pts)

(iii) *Assumptions.* For the t-test, we skipped the step where we check assumptions of the test. Using your histogram from part (a), which test should we have run? Justify your choice. (3 pts)