STAT 324 Spring 2020: Homework 2 Solutions + Rubric

To receive credit, you must submit your assignment to Canvas before **6pm, Thursday, February 6**. The file submission must be a knitted .html or .pdf file, made using RMarkdown. The code you used to answer the questions should be included in your file. You do not need to submit your .rmd file. The assignment is worth 50 points.

##50 Total Points

##Questions:

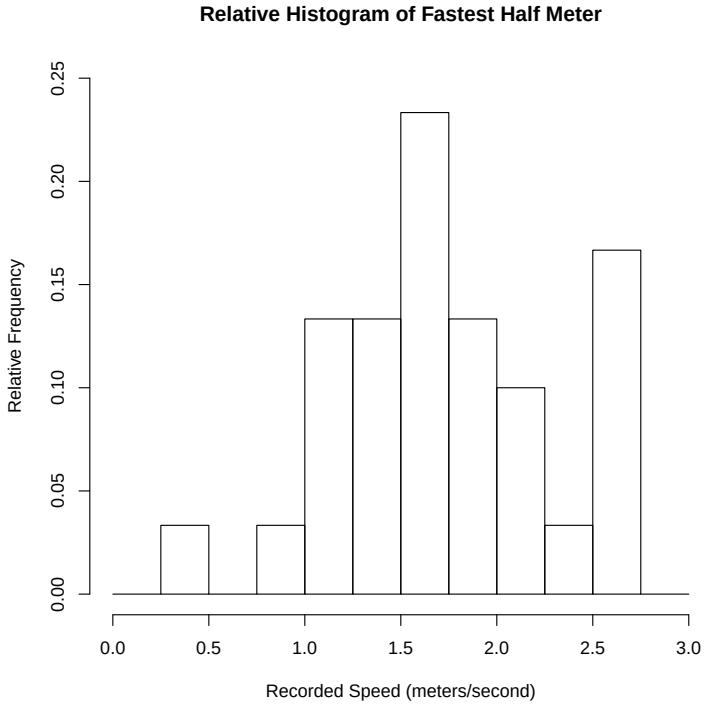
1. There are 12 numbers on a list, and the mean is 24. The smallest number on the list is changed from 11.9 to 1.19.

   (a) Is it possible to determine the direction in which (increase/decrease) the mean changes? Or how much the mean changes? If so, by how much does it change? If not, why not? ***3 points +1 yes, +1 decrease, +1 for amount of change*** *Yes. Original mean is 24, which means original total is* $12 * 24 = 288$. *New total is* $288 - 11.9 + 1.19 = 277.29$. *New mean is* $277.29/12 = 23.1075$, *so mean decreased by* $24 - 23.1075 = 0.8925$

   (b) Is it possible to determine the direction in which the median changes? Or how much the median changes? If so, by how much does it change? If not, why not? ***3 points +1 yes, +2 for no change****Yes. Since median is related to a middle value and the smallest number is not a middle number in a list of 12, so the median will not be affected. Change=0.*

   (c) Is it possible to predict the direction in which the standard deviation changes? If so, does it get larger or smaller? If not, why not? Describe why it is difficult to predict by how much the standard deviation will change in this case. ***2 points, +1 for yes, +1 gets larger*** *Yes, our new value is much lower than the original minimum which spreads out our data and increases the standard deviation of the set (it may be a good idea to try out a few examples to show this to yourself). We do not know the original standard deviation or original values however, so we're unable to calculate the change in standard deviation.*

2. A zoologist collected wild lizards in the Southwestern United States. Thirty lizards from the genus *Phrynosoma* were placed on a treatmill and their speed measured. The recorded speeds (meters/second) (the fastest time to run a half meter) for the thirty lizards are summarized in the relative histogram below. (Data Courtesy of K. Bonine *)



1pt each

(a) Is the percent of lizards with recorded speed below 1.25 closest to: 25%, 50%, or 75%? *25%*

(b) In which interval are there more speeds recorded: 1.5-1.75 or 2-2.5? *1.5-1.75 looks to have about 23%, while 2-2.5 has about 13%*

(c) About how many lizards had recorded speeds above 1 meters/second? *28 since it looks like there were 2 below 1 meters/second*

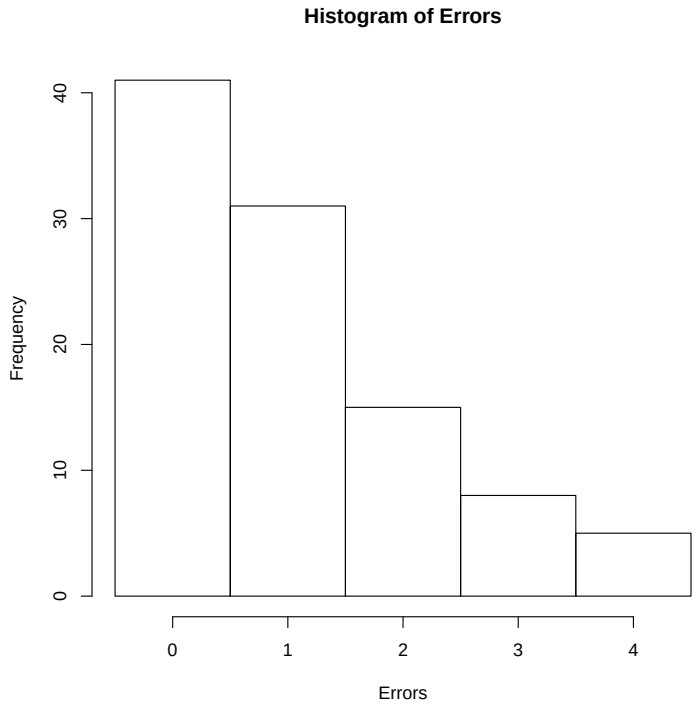
(d) In which bin does the median fall? Show how you know. *we are looking for the bin with 50% below and 50% above. So, lets add up until we get to 50%: 3.33+3.33+13.33+13.33=33.32 and adding the next bin: +23.33=56.65, so the median falls in the 1.5-1.75 bin.*

3. After manufacture, computer disks are tested for errors. The table below gives the number of errors detected on a random sample of 100 disks. Hint: You can use the rep() function in R to make a vector of repeated numbers.

Number of Defects	0	1	2	3	4
Frequency	41	31	15	8	5

(a) Describe the type of data (ex: nominal) that is being recorded about the sample of 100 disks, being as specific as possible. ***2 points +1 Numeric/Quantitative, +1 Discrete***

(b) Construct a frequency histogram of the information with R. **3 points**



(c) What is the shape of the histogram for the number of defects observed in this sample?
1 point *Right Skewed*

(d) Calculate the mean and median number of errors detected on the 100 disks by hand and with R. How do the mean and median values compare and is that consistent with what we would guess based on the shape?
3 point +1 for each value, +1 comparison *mean: 1.05, median: 1. The mean is slightly above the median which is consistent with the right skew. Median is closer to higher frequency lower values*

(e) Calculate the sample standard deviation with R. Explain what this value means in the context of the problem.
2 point, +1 for values, +1 explanation *sd= 1.157976. The number of defaults on a computer disk in the sample of 100 were about 1.16 deviations away from the mean of 1.05 errors, on average.*

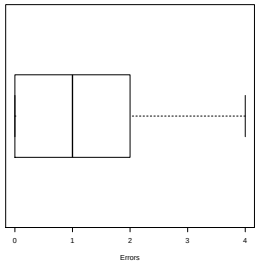
(f) Calculate the first and third quartiles and IQR by hand and with R. Are the values consistent betweeen the two methods? Explain what the three values mean in the context of the problem.
3 points +1 for values by hand, +1 for values on R, +1 for comparison (same) *Since there are 100 data values, Median is the average of the 50 and 51st values. Median=1. Q1 is then the median of the first 50 observations, so the average of 25th and 26th observations, which are both 0: Q1=0. Q3 is the median of the last 50 observations, so the average of the 75th and 76th observations which are both 2, so Q3 is 2. These are the same values calculated by R. The IQR is 2-0=2. Q1 tells us the number of errors below which approximately 25% of the disks experienced. Q3 tells us the number of errors below which approximately 75% of the disks experienced. IQR tells us the range of the middle 50% of errors.*

(g) What proportion of the computer disks had a number of errors greater than the mean number of

errors?
1 point *Since mean was 1.05, those with 2, 3 or 4 had a number greater than the mean so that is 15+8+5=28/100 or 28% of the disks*

(h) What range of values for this sample data are not considered outliers using the [Q1-1.5IQR, Q3+1.5IQR] designation (using the IQR you calculated by hand)?.
1 point *1.5*IQR=3, so Q1-1.5*IQR=-3, Q3+1.5*IQR=5. Any values between 0 and 5 are not considered outliers for this sample. Saying -3 is ok.*

(i) Make a boxplot of the data using R and compare the lines to the values you calculated by hand.
2 points. +1 for plot, +1 for comparison

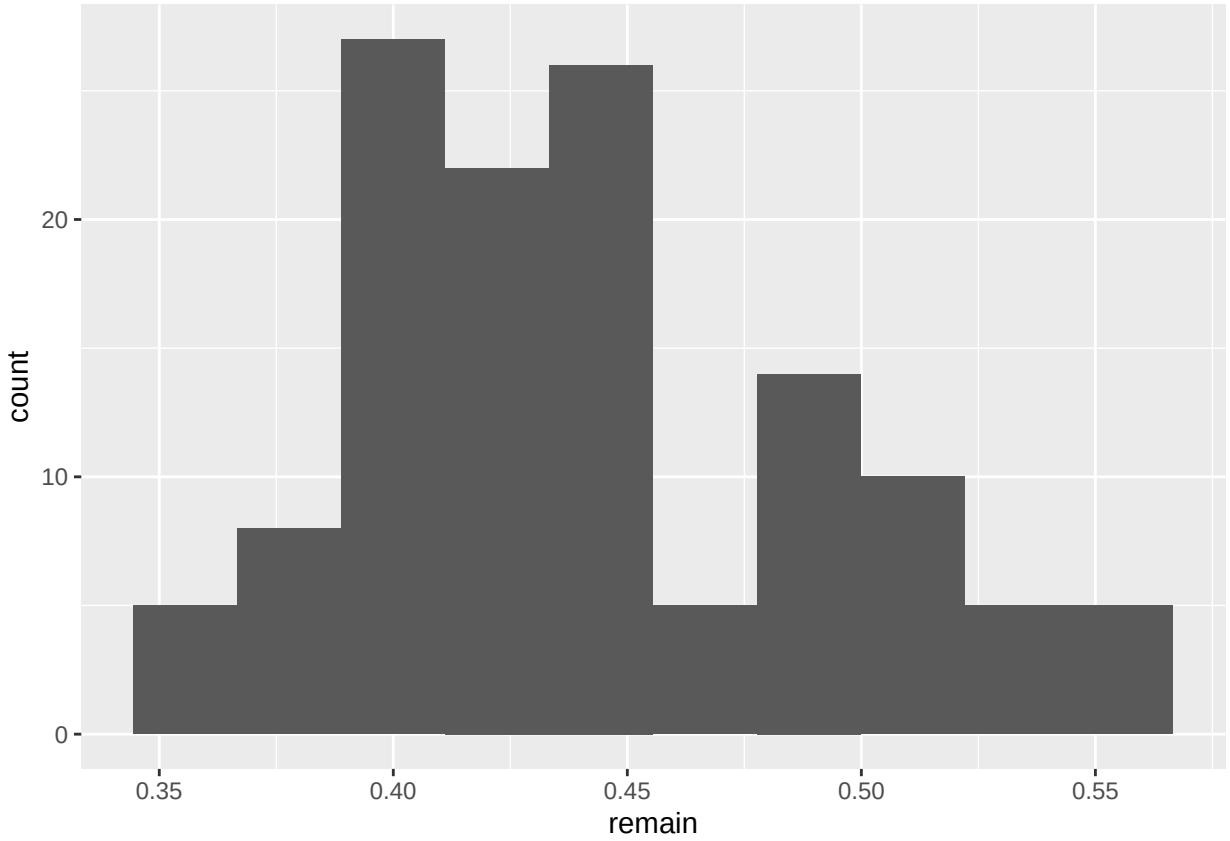


(j) Compare and contrast (briefly) the information about the data given by the histogram in part b and the boxplot in part i. ***1pt*** *Both graphs show the right skew; The histogram more clearly shows the discrete nature of the data; the boxplot more clearly identifies the 5 number summary (Min, Q1, Med, Q3, Max)*

5. The file brexit.csv contains the results of 127 polls, including both online polls and telephone polls, carried out from January 2016 to the referendum date on June 23, 2016. Use that dataset to answer the following questions.

   a. Use R to create a histogram for the proportion who answered "Remain" when polled. Describe the shape of the data.

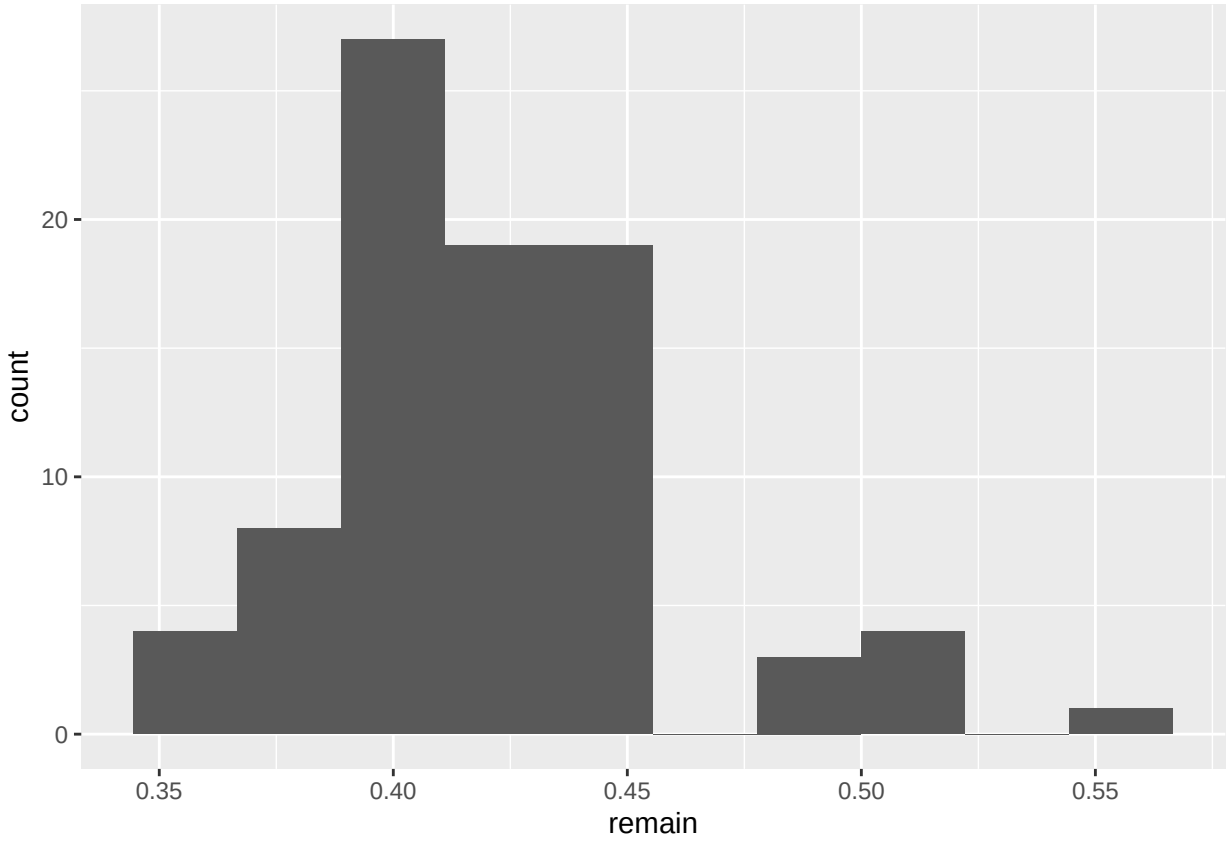
```
library(tidyverse)
brexit <- read_csv("brexit.csv")
ggplot(data=brexit,aes(x=remain)) +
  geom_histogram(bins = 10)
```



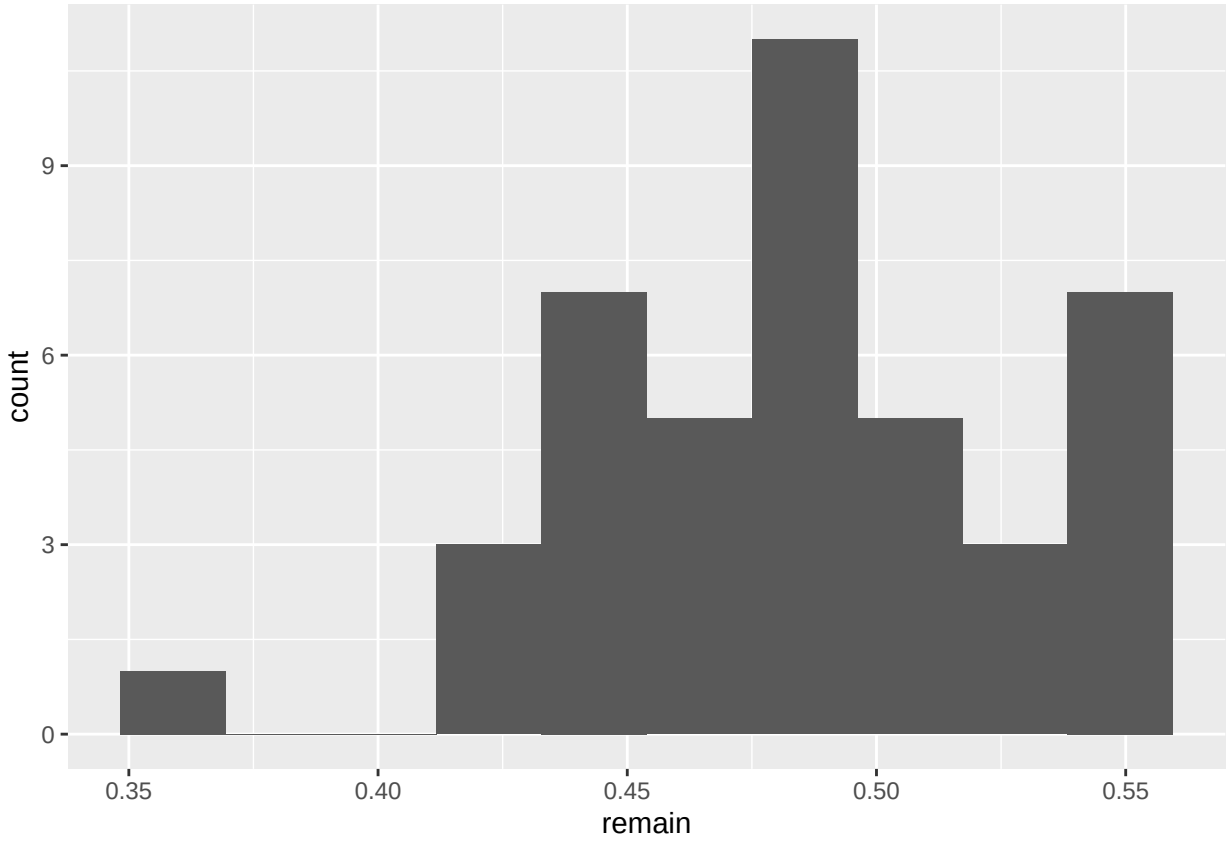
(2pts, +1 for plot, +1 for mentioning bimodality or other reasonable shape) *The data looks somewhat bimodal, or perhaps skewed right.*

b. Now construct two separate histograms for the proportion who answered "Remain". Make one histogram for online polls and another histogram for telephone polls. Describe the shape and relative position of the data.

```
online <- brexit %>% filter(poll_type=="Online")
phone <- brexit %>% filter(poll_type=="Telephone")
ggplot(data=online,aes(x=remain)) +
  geom_histogram(bins=10)
```




```
ggplot(data=phone,aes(x=remain)) +
  geom_histogram(bins=10)
```



(3pts, +1 for each plot, +1 for correct comparison)*Both sets of data are roughly symmetric, the online poll data has a very slight right skew, while the telephone poll data has a very slight left. The online poll results are positioned left of the telephone poll results, a smaller proportion of people answered "Remain" online compared to over the phone.*

c. Compute the mean and median proportion voting "Remain" observed for the online and telephone polls. Compare both measures of center across the two groups.

```
brexit %>%
  group_by(poll_type) %>%
  summarize(median_remain=median(remain),mean_remain=mean(remain))
```

```
## # A tibble: 2 x 3
##   poll_type median_remain mean_remain
##   <chr>             <dbl>       <dbl>
## 1 Online             0.42       0.421
## 2 Telephone          0.48       0.485
```

(2pts, +1 for mean, +1 for median)*In terms of both mean and median, about 6% more people answered "Remain" over the phone compared to online.*

d. Compute and compare the standard deviation observed in the two groups.

```r
sd(online$remain)
```

```
## [1] 0.03750518
```

```r
sd(phone$remain)
```

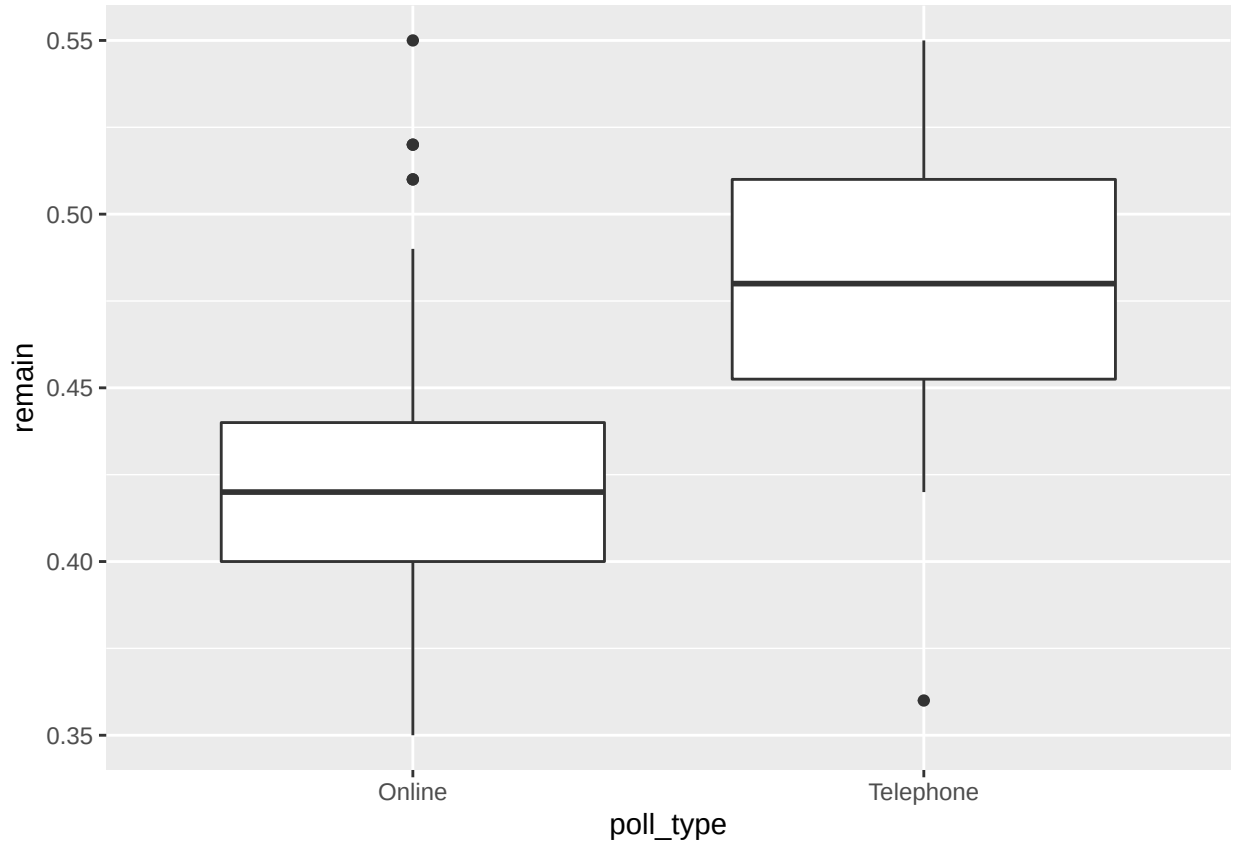
```
## [1] 0.04232568
```

(1pt) *The standard deviations are similar for both groups, at around 4% for each. The standard deviation for phone results was slightly greater.*

e. Use R to help you create side by side boxplots of the two sets so they are easily comparable.

(2pts for a correct, labeled plot)

```r
ggplot(data=brexit,aes(y=remain,group=poll_type,x=poll_type)) +
  geom_boxplot()
```



f. How many values were identified as outliers? Would these values have been identified as an outlier in the other type of poll? Use the 1.5IQR rule for identifying outliers.

```r
max(online$remain)
```

```
## [1] 0.55
```

```r
summary(phone$remain)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3600  0.4525  0.4800  0.4850  0.5100  0.5500
```

```r
0.51+(0.51-0.4525)
```

```
## [1] 0.5675
```

(2pts, +1 for recognizing the 4 outliers, +1 for recognizing none would be outliers in other group)*4 values were identified as outliers, 3 in the online polls and 1 in the telephone polls. None of them would be considered outliers in the other type of poll.*

h. What would be the mean and median proportion answering "Remain" if we combined the two poll types together? Show how one of these can be calculated directly from your summary measures in part (c).

```r
median(brexit$remain)
```

```
## [1] 0.44
```

(4 pts: +1 combined mean, +1 combined median, +2 combined mean calculation) *mean: 0.442 is the weighted average of the means found in part c:* $\frac{850.422+420.485}{127} = 0.442$*. There is no good way to get the combined median from the above medians. R finds a combined median of 0.44*

i. Next, calculate the mean proportion of respondents that answered "Leave" for both online and telephone polls. What other factor in the data can explain the much smaller gap between means here compared to part c? Explain.

```r
brexit %>% group_by(poll_type) %>% summarize(mean_leave=mean(leave))
```

```
## # A tibble: 2 x 2
##   poll_type mean_leave
##   <chr>          <dbl>
## 1 Online         0.426
## 2 Telephone      0.415
```

(4pts,+2 for mean calculation, +2 for explanation)*The difference between means in part c was over 0.06, while the difference between means here is about 0.01. This can be explained by the fact that more people responded "Undecided" in online polls compared to phone polls. In other words, about the same proportion answered "Leave" in both types of polls, but more people answered "Undecided" in online polls than phone polls, so fewer people could answer "Remain" in online polls.*

```
ggplot(data=brexit,aes(y=undecided,group=poll_type,x=poll_type)) +
  geom_boxplot()
```