Stat 371 Homework #2

Student's Name Here

- Submit your homework to Canvas by the due date and time. Email your lecturer if you have extenuating circumstances and need to request an extension.
- If an exercise asks you to use R, include a copy of all relevant code and output in your submitted homework file. You can copy/paste your code, take screenshots, or compile your work in an Rmarkdown document.
- If a problem does not specify how to compute the answer, you many use any appropriate method. I may ask you to use R or use manual calculations on your exams, so practice accordingly.
- You must include an explanation and/or intermediate calculations for an exercise to be complete.
- Be sure to submit the HWK2 Autograde Quiz which will give you ~20 of your 40 accuracy points.
- 50 points total: 40 points accuracy, and 10 points completion

Summarizing Data Numerically and Graphically

Exercise 1. There are $n = 12$ numbers in a sample, and the mean is $\bar{x} = 24$. The minimum of the sample is accidentally changed from 11.9 to 1.19.

a. Is it possible to determine the direction (increase/decrease) in which the mean $\bar{x}$ changes? And how much the mean changes? If so, by how much does it change? If not, why not?

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?

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.

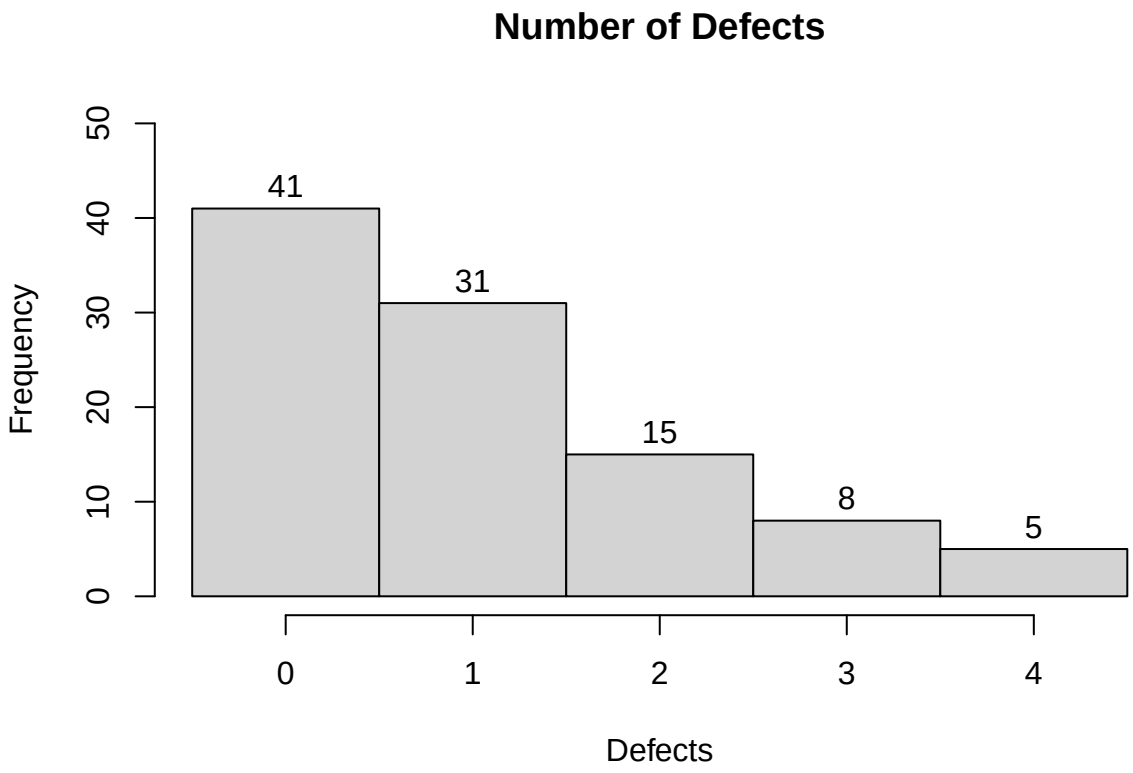
Exercise 2. Recall the computer disk error data given used in Homework 1. The table below tabulates the number of errors detected on each of the 100 disks produced in a day.

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

A frequency histogram (without the typo from Homework 1) showing the frequency for number of errors on the 100 disks is given below.

```r
error.data <- c(rep(0,41), rep(1,31), rep(2,15), rep(3,8), rep(4, 5))

hist(error.data, breaks=c(seq(from = -0.5, 4.5, by = 1)),
     xlab = "Defects", main = "Number of Defects", labels = TRUE,
     ylim = c(0, 50))
```



a. What is the shape of the histogram for the number of defects observed in this sample? Why does that make sense in the context of the question?

b. Calculate the mean and median number of errors detected on the 100 disks 'by hand' and using the built-in R functions. How do the mean and median values compare and is that consistent with what we would guess based on the shape?

c. Calculate the sample standard deviation "by hand" and using the built in R function. Are the values consistent between the two methods?

d. Construct a boxplot for the number of errors data using R with helpful labels. Explain how the shape of the data identified in (a) can be seen from the boxplot.

e. Explain why the histogram is better able to show the discrete nature of the data than a boxplot.

Exercise 3. A certain reaction was run several times using each of two catalysts, A and B. The catalysts are supposed to control the yield of an undesireable side product. Results, in units of percentage yield, for 25 runs of catalyst A and 23 runs of catalyst B are given below and also in `Catalysts.csv`.

Catalyst A: 4.3, 4.4, 3.4, 2.6, 3.8, 4.9, 4.6, 5.2, 4.7, 4.1, 2.6, 6.7, 4.1, 3.6, 2.9, 2.6, 4.0, 4.3, 3.9, 4.8, 4.5, 4.4, 3.1, 5.7, 4.5

Catalyst B: 3.4, 5.9, 1.2, 2.1, 5.5, 6.4, 5.0, 5.8, 2.5, 3.7, 3.8, 5.1, 3.1, 1.6, 3.5, 5.9, 6.7, 5.2, 5.8, 2.2, 4.3, 3.8, 1.2

a. Use R to create a histogram for the percentage yield of the undesireable side product for the two catalysts (any kind of histogram that you want since sample sizes are similar). Have identical x and y axis scales so the two groups' values are more easily compared. Include useful titles.

b. Compare the shape of the percentage yields from the two catalysts observed in this sample.

c. Compute the mean and median percentage yields observed for catalyst A and catalyst B using R. Compare both measures of center within each group and comment on how that relationship corresponds to the datas' shapes. Also compare the measures of center across the two groups and comment on how that relationship is evident in the histograms.

d. Compute (in R) and compare the sample standard deviation of percentage yield from catalyst A and catalyst B. Comment on how the relative size of these values can be identified from the histograms. Describe in words what these values mean when considering which catalyst to use for your experiment.

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

f. Explain why the highest value in the catalyst A boxplot is shown as a point. That is, explain what calculations determined that 6.7 was an outlying value. Also specify to what value the upper catalyst A percentage yield whisker extends.

g. What would be the mean and median percentage yield if we combined the two data sets into one large data set? Show how the mean can be calculated from the summary measures in part (c) along with the sample sizes and explain why the median of the combined set cannot be computed based on (c).

Exercise 4. You are adding Badger-themed bedazzle to your striped overalls and are using both red and white beads. You are interested in how the size of the bag of beads you select your beads from changes the probability of outcomes of interest. Compute the probability for the outcomes in (a) and (b) for all three different sampling strategies.

(Small Pop): Drawing without replacement from a small bag of beads with 7 White beads and 3 Red beads.

(Large Pop): Drawing without replacement from a large bag of beads with 700 White beads and 300 Red beads.

(Same Pop): Drawing from a bag of beads that always contains exactly 70% White and 30% Red beads.

For example, consider choosing 3 beads. Calculate the probability of selecting no white beads.

Small Pop: $\text{P(RRR)} = \frac{3}{10} * \frac{2}{9} * \frac{1}{8} = 0.008333333$

Large Pop: $\text{P(RRR)} = \frac{300}{1000} * \frac{299}{999} * \frac{298}{998} = 0.02681098$

Same Pop: $\text{P(RRR)} = 0.30 * 0.30 * 0.30 = 0.027$

a. Consider choosing 3 beads. Calculate the probability of selecting exactly 1 white bead.

b. Consider choosing 3 beads. Calculate the probability of selecting at least 1 white bead.

c. Consider sampling without replacement. Does drawing from a population that is **small** or **large** relative to your sample size result in an probability that is closest to the probability when sampling with replacement?