Stat 324 Homework #2

Svadrut Kukunooru

*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 the code and output. Please edit your code and output to be only the relevant portions.

*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

Basics of Statistics and Summarizing Data Numerically and Graphically (I)

Exercise 1. There are 12 numbers in a sample, and the mean is $\bar{x} = 27$. The minimum of the sample is accidentally changed from 13.8 to 1.38.

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

The mean decreases because the minimum becomes much smaller. We can calculate how much it changes by calculating the total sum of all the elements by multiplying the mean by the number of elements:

$$27 \times 12 = 324$$

Calculate the difference in change:

$$13.8 - 1.38 = 12.42$$

Then, calculate the mean with the new sum:

$$324 - 12.42 = 311.58311.58 \div 12 = \boxed{25.965}$$

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? How do you know?

The median does not change, as it only depends on the position of the elements, not their value. Since the minimum simply gets smaller, this does not have any effect on the value of the median. However, if the minimum was removed, the median would then change because there would be a different number of elements.

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? How do you know?

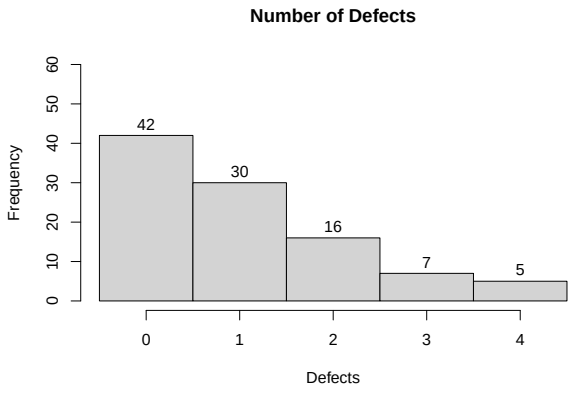
Since we are essentially increasing the spread of elements by decreasing the minimum (therefore making the difference between the minimum and the mean bigger), the standard deviation will get larger.

Exercise 2: Recall the computer disk error data given in HWK 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	42
1	30
2	16
3	7
4	5

A frequency histogram showing the frequency for number of errors on the 100 disks is given below.

```
error.data=c(rep(0,42), rep(1,30), rep(2,16), rep(3,7), 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(0,60))
```



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?

The shape of the histogram is decreasing exponentially. This makes sense in the context of the question because it would make sense that there would be a lesser and lesser chance for there to be errors on a disk as the number of errors increase.

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? [You can use the text such as $\bar{x} = \frac{value1}{value2}$ to help you show your work neatly].

MEAN BY HAND:
$(42 \times 0) + (30 \times 1) + (16 \times 2) + (7 \times 3) + (5 \times 4)$
$= 0 + 30 + 32 + 21 + 20 = 103$
$103/100 = \boxed{1.03}$

MEDIAN BY HAND:
Since there is 100 disks, the median disk will be between 50 and 51. Therefore, the median is $\boxed{1}$.

```
print(mean(error.data))
```

```
## [1] 1.03
```

```
print(median(error.data))
```

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

The values are the same. This is consistent based on what we can guess with the shape, as there are much more disks with 0 or 1 errors than more errors.

c. Calculate the sample standard deviation "by hand" and using the built in R function. Are the values consistent between the two methods? How would our calculation differ if instead we considered these 100 values the whole population? hint: use multiplication instead of repeated addition

$(42 \times 1.03^2) + (30 \times 0.03^2) + (16 \times 0.97^2) + (7 \times 1.97^2) + (5 \times 2.97^2) = 130.91$
$130.91/100 = 1.3091$
$\sqrt{1.3091} = \boxed{1.144159}$

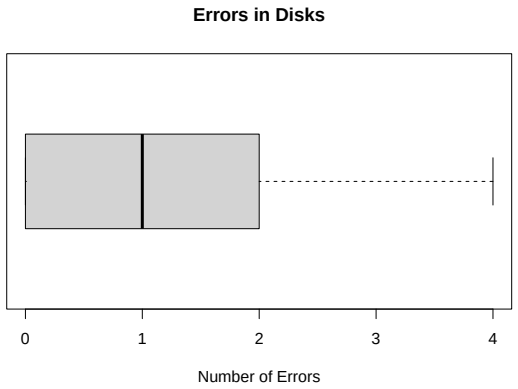
```
print(sd(error.data))
```

```
## [1] 1.149923
```

The values are not consistent between the two methods since calculating it by hand may have an error and also is not as precise as using the computer version.

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.

```
boxplot(error.data, pch=20, whisker.width=0.5, main="Errors in Disks", horizontal = TRUE, xlab = "Numbe
```



The shape of the data can be seen in the boxplot because because all the data from the minimum to the upper quartile is in the range 0-2, reflecting that most of the disks also have this range of errors.

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

The histogram is better able to show the discrete nature of the data because it shows how many disks have each number of errors, unlike a boxplot which only shows the quartiles and the minmax.

Exercise 3: A company that manufactures toilets claims that its new presure-assisted toilet reduces the average amount of water used by more thaan 0.5 gallons per flush when compared to its current model. The company selects 20 toilets of the current type and 19 of the New type and measures the amount of water used when each toilet is flushed once. The number of gallons measured for each flush are recorded below. The measurements are also given in flush.csv.

Current Model: 1.63, 1.25, 1.23, 1.49, 2.11, 1.48, 1.94, 1.72, 1.85, 1.54, 1.67, 1.76, 1.46, 1.32, 1.23, 1.67, 1.74, 1.63, 1.25, 1.56

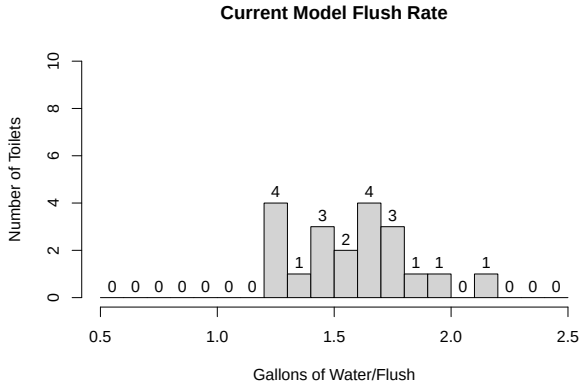
New Model: 1.28, 1.19, 0.90, 1.24, 1.00, 0.80, 0.71, 1.03, 1.27, 1.14, 1.36, 0.91, 1.09, 1.36, 0.91, 0.91, 0.86, 0.93, 1.36

a. Use R to create histograms to display the sample data from each model (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.

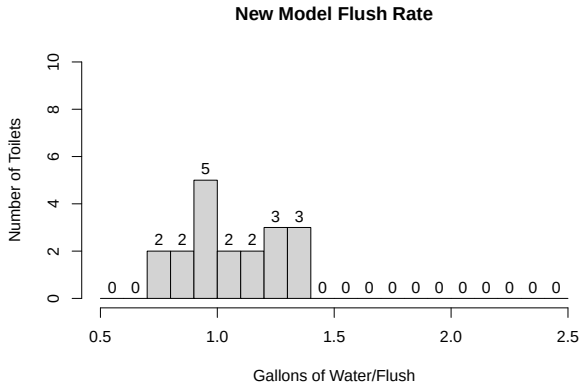
```
currentModel = c(1.63, 1.25, 1.23, 1.49, 2.11, 1.48, 1.94, 1.72, 1.85, 1.54, 1.67, 1.76, 1.46, 1.32, 1.

newModel = c(1.28, 1.19, 0.90, 1.24, 1.00, 0.80, 0.71, 1.03, 1.27, 1.14, 1.36, 0.91, 1.09, 1.36, 0.91, 


hist(currentModel,
     breaks=c(seq(from=0.5, 2.5, by=0.1)),
     xlab="Gallons of Water/Flush", ylab="Number of Toilets", main="Current Model Flush Rate",
     labels=TRUE, ylim=c(0,10))
```




```
hist(newModel,
     breaks=c(seq(from=0.5, 2.5, by=0.1)),
     xlab="Gallons of Water/Flush", ylab="Number of Toilets", main="New Model Flush Rate",
     labels=TRUE, ylim=c(0,10))
```



b. Compare the shape of the gallons flushed from the two models of toilets samples.

It appears that the new model flushes much less on average than the current model does.

c. Compute the mean and median gallons flushed for the Current and New Model toilets using the built-in R function. 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.

```
mean(currentModel)
```

```
## [1] 1.5765
```

```
median(currentModel)
```

```
## [1] 1.595
```

```r
mean(newModel)
```

```
## [1] 1.065789
```

```r
median(newModel)
```

```
## [1] 1.03
```

The median and mean in the current model are much higher than the median and mean in the new model, showing that on average the new model spends less gallons of water per flush.

d. Compute (using built-in R function) and compare the sample standard deviation of gallons flushed by the current and new model toilets. Comment on how the relative size of these values can be identified from the histograms.

```r
sd(currentModel)
```

```
## [1] 0.2456843
```

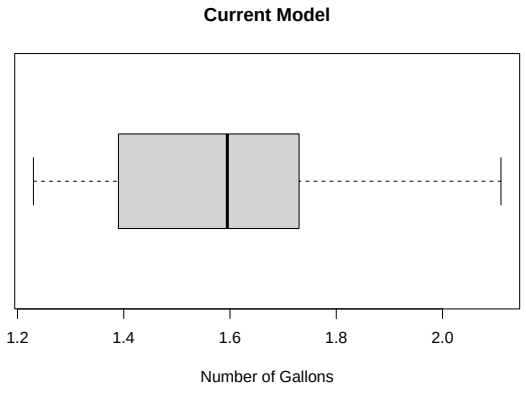
```r
sd(newModel)
```

```
## [1] 0.2058941
```

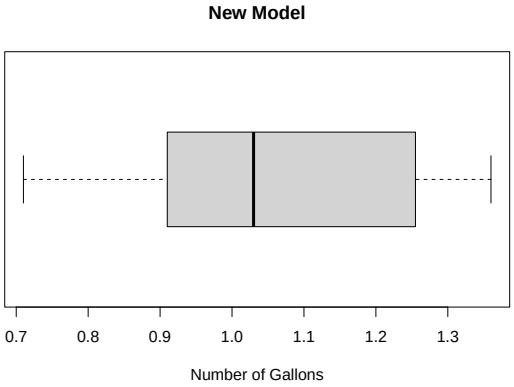
The new model has a lower standard deviation, and this is seen by the lesser spread of the data in the histogram of the new model compared to the current model.

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

```r
currentBox = boxplot(currentModel, pch=20, whisker.width=0.5, main="Current Model", horizontal = TRUE, x
```




```r
newBox = boxplot(newModel, pch=20, whisker.width=0.5, main="New Model", horizontal = TRUE, xlab = "Numbe
```



f. Explain why there are no values shown as a dot (outlier) on the Current Model flush boxplot. To what values do the Current model flush boxplot whiskers extend? (Use R for your boxplot calculations and type=2 for quantiles)

```
currentBox$stats
```

```
##       [,1]
## [1,] 1.230
## [2,] 1.390
## [3,] 1.595
## [4,] 1.730
## [5,] 2.110
```

There are no values shown as a dot on the current model flush boxplot because there are no values outside the 1.5 * IQR of the data. The whiskers extend to 1.230 and 2.110.

g. What would be the mean and median gallons flushed if we combined the two data sets into one large data set with 39 observations? Show how the mean can be calculated using R and then from the summary measures in part (c) along with the sample sizes. Explain why the median of the combined set cannot be computed based on the summaries in part (c).

```
newData = c(currentModel, newModel)
mean(newData)
```

```
## [1] 1.327692
```

```
median(newData)
```

```
## [1] 1.28
```

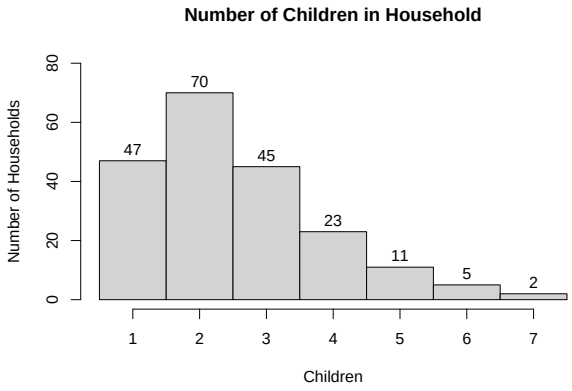
```
# Now calculate the new mean by adding the two means from part (c) and dividing by 2.
(1.5765 * 20 + 1.065789 * 19)/39
```

```
## [1] 1.327692
```

The median cannot be calculated from the results in part(c) since the median will change based on each of the values in each set. We cannot simply perform a calculation on the two medians.

Exercise 4: An elementary school surveys its families and tabulates the number of children reported in each household. A frequency histogram summarizes the data received:

```
Children=c(rep(1, 47), rep(2, 70), rep(3, 45), rep(4, 23), rep(5,11), rep(6,5), rep(7,2))
hist(Children,
     breaks=seq(0.5, 7.5, 1),
     labels=TRUE,
     ylim=c(0,80),
     main="Number of Children in Household",
     ylab="Number of Households")
```



a. Consider a randomly chosen household, Household A. Identify whether the events "Household A has 1 Child" and "Household A has 2 Children" are (i) independent, (ii) mutually exclusive, (iii) both independent and mutually exclusive or (iv) neither independent nor mutually exclusive. Explain how you know.

These two events are neither independent nor mutually exclusive. Household A cannot have two children and one child at the same time. Additionally, if a household has 1 child, it is more likely for the household to have two children than no children. Therefore, these two events are dependent.

b. Suppose the principal chooses a random family from those at the school to call each day and each family is equally likely to be chosen on the first day of school. What is the probability that the family has more than two (2) children?

0.4236453

c. Suppose the principal randomly chooses a family to call from those at the school that they have not already called. What is the probability that all of the families called the first 5 days of school had a single (1) child in the household? Is this a highly likely or unlikely event?

The probability is 0.0005611291. This is a highly unlikely event.