HW1

Teryl Schmidt

9/15/2018

Problem 1

If you wanted to estimate the mean height of all the students at UW Madison, which one of the following sampling strategies would be best? Why? Note that none of the methods are true simple random samples.

a. Measure the heights of 100 students found in the gym during basketball intramurals.
b. Measure the heights of the engineering majors.
c. Measure the heights of the students selected by choosing the first name on each page of a list of students enrolled that semester.

Problem 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 *)

a. Is the percent of lizards with recorded speed below 1.25 closest to: 25% , 50%, or 75%?
b. In which interval are there more speeds recorded: 1.5-1.75 or 2-2.5?
c. About how many lizards had recorded speeds above 1 meters/second?

~0.04 + ~0.04 = 0.08 below 1 m/s

~0.92 above 1 m/s

92% of 30 lizards is 27.6 lizards

~28 lizards have recorded speeds above 1 m/s

Problem 3

In a sample of 30 men, the mean height was 179 cm with standard deviation of 6 cm. In a sample of 25 women, the mean height was 163 cm with standard deviation of 6 cm. If both samples were combined into one larger group…

a. What is the mean height for the combined group?

~157.4167

```
((25 * 163) + (30 * 179)) / (25 + 35)
```

```
## [1] 157.4167
```

b. The standard deviation for the combined group would be

    i. Less than 6 cm
    ii. Greater than 6 cm
    iii. Equal to 6 cm
    iv. Can not tell from the information given

```
((30 * (6^2 + (179 - 157.4167))) + (25 * (6^2 + (163 - 157.4167)))) / (25 + 30)
```

```
## [1] 50.31057
```

```
sqrt(50.31057)
```

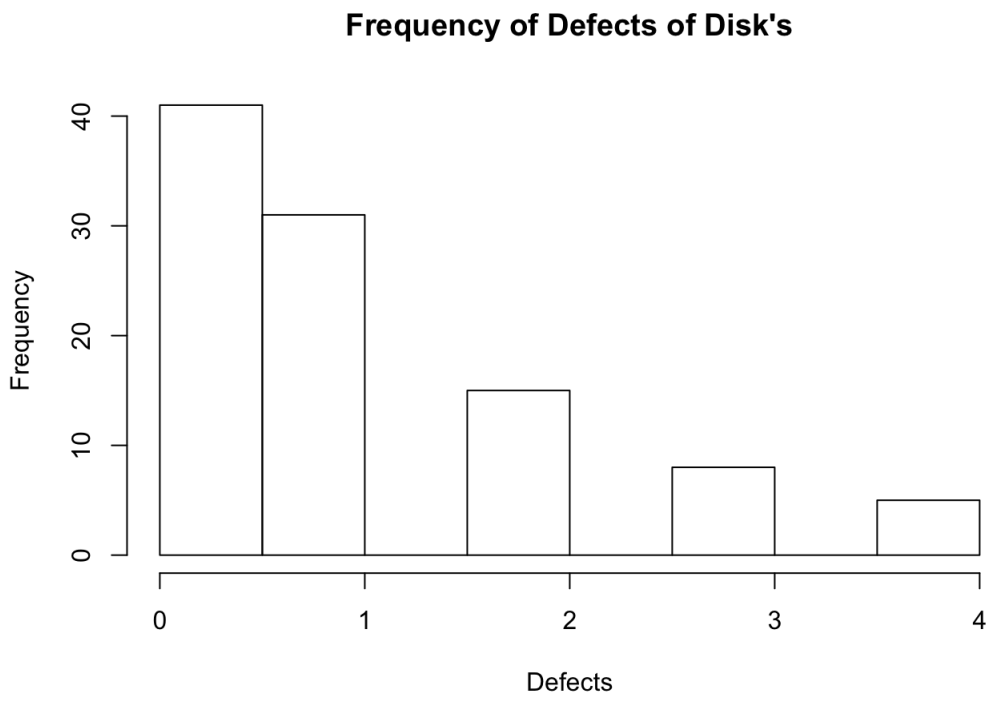
```
## [1] 7.092994
```

Problem 4

After manufacture, computer disks are tested for errors. The table below gives the number of errors detected on a random sample of 100 disks.

a. Construct a frequency histogram of the information by creating bins at [-.5, .5), [.5, 1.5), etc.

```
Defects = c(0,1,2,3,4)
Frequency = c(rep(0, 41), rep(1, 31), rep(2, 15), rep(3, 8), rep(4, 5))
hist(Frequency, main="Frequency of Defects of Disk's", xlab="Defects")
```



b. What is the shape of the histogram for the number of defects observed in this sample?

Right Skewed

c. Calculate the mean and median number of errors detected on the 100 disks. How do these values compare and is that consistent with what we would guess based on the shape?

```
mean(Frequency)
```

```
## [1] 1.05
```

```
median(Frequency)
```

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

The mean is more than the median so the histograph is skewed to the right.

d. Calculate the sample standard deviation with your calculator and R. Are the values consistent between the two methods? Explain what this value means in the context of the problem.

```
sd(Frequency)
```

```
## [1] 1.157976
```

```
mean = mean(Frequency)

for (value in Frequency) {
  value = (value - mean)^2
}

sqrt(mean(Frequency))
```

```
## [1] 1.024695
```

Not the exact same number due to rounding. This value means how far or an average value is from the mean.

e. 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.

```
IQR(Frequency)
```

```
## [1] 2
```

```
quantile(Frequency, c(0.25, 0.75))
```

```
## 25% 75%
##   0   2
```

```
IQR = quantile(Frequency, prob=0.75)-quantile(Frequency, prob=0.25)
IQR
```

```
## 75%
##   2
```

The IQR is how spread out the middle values are. The spread goes from 0 to 2, which means a lot of the data is around 0 to 2, and there are not many data points from 2 to 4.

f. What proportion of the computer disks had a number of errors greater than the mean number of errors?

```
(15 + 8 + 5) / 100
```

```
## [1] 0.28
```

~28%

g. 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)?.

```
0 - (1.5 * IQR)
```

```
## 75%
##  -3
```

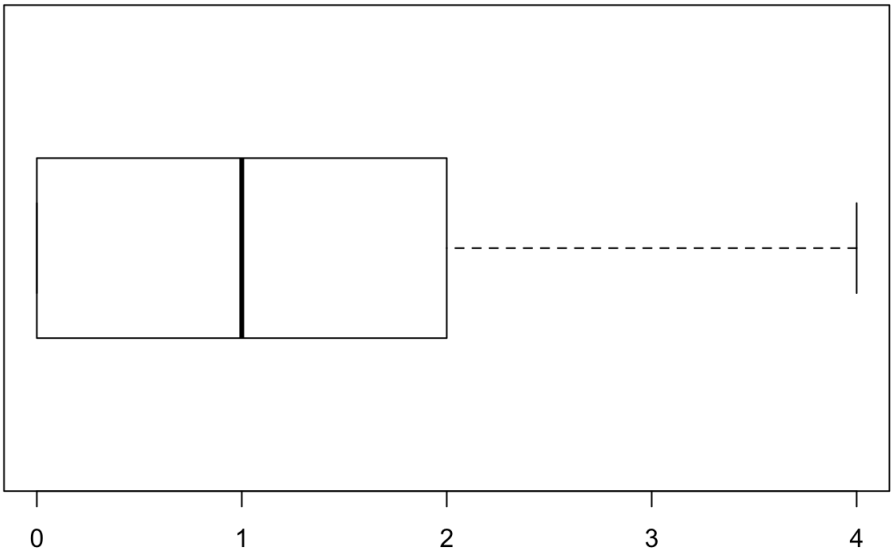
```
2 + (1.5 * IQR)
```

```
## 75%
##   5
```

Outliar range = x < -3 , 5 < x

h. Sketch a boxplot of the data by hand (using the relevant values you calculated by hand).

```
boxplot(Frequency, horizontal = T)
```



i. Compare and contrast (briefly) the information about the data given by the histogram in part a and the boxplot in part h.

A Boxplot shows the minimum value, the first and third quartile, and the mean value. The histogram displays the frequencies of values.

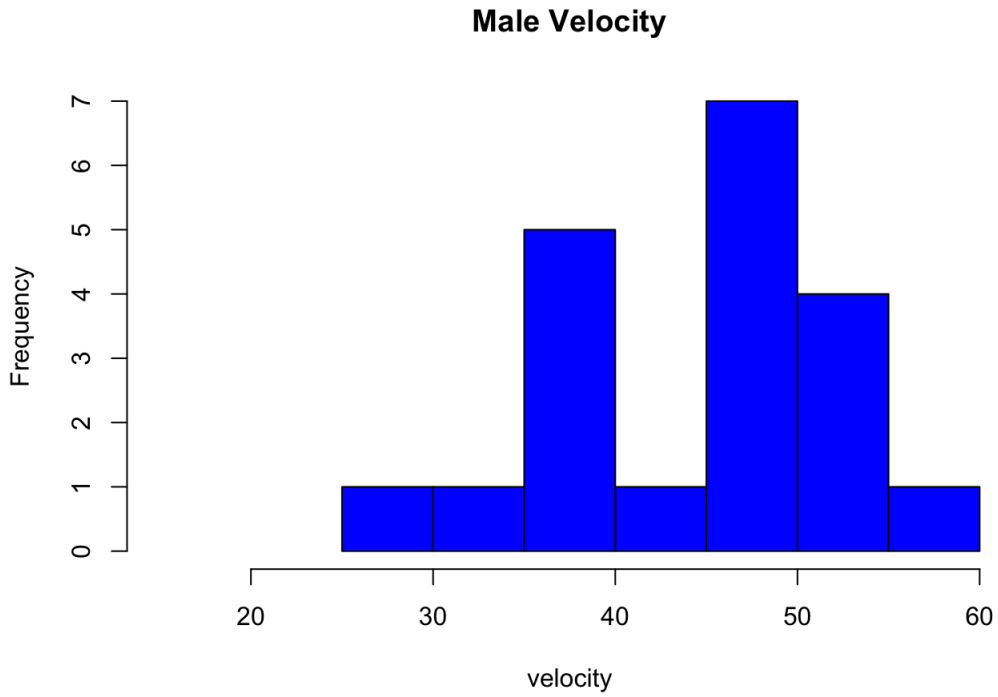
Problem 5

Physical education researchers interested in the development of the overarm throw measured the horizontal velocity of a thrown ball at the time of release. The results for first-grade children (in feet/sec) (courtesy of L. Halverson and M. Roberton*) are: Males: 54.2, 39.6, 52.3, 48.4, 35.9, 30.4, 25.2, 45.4, 48.9, 48.9, 45.8, 44.0, 52.5, 48.3, 59.9, 51.7, 38.6, 39.1, 49.9, 38.3 Females: 30.3, 43.0, 25.7, 26.7, 27.3, 31.9, 53.7, 32.9, 19.4, 23.7, 23.3, 23.3, 37.8, 39.5, 33.5, 30.4, 28.5 3

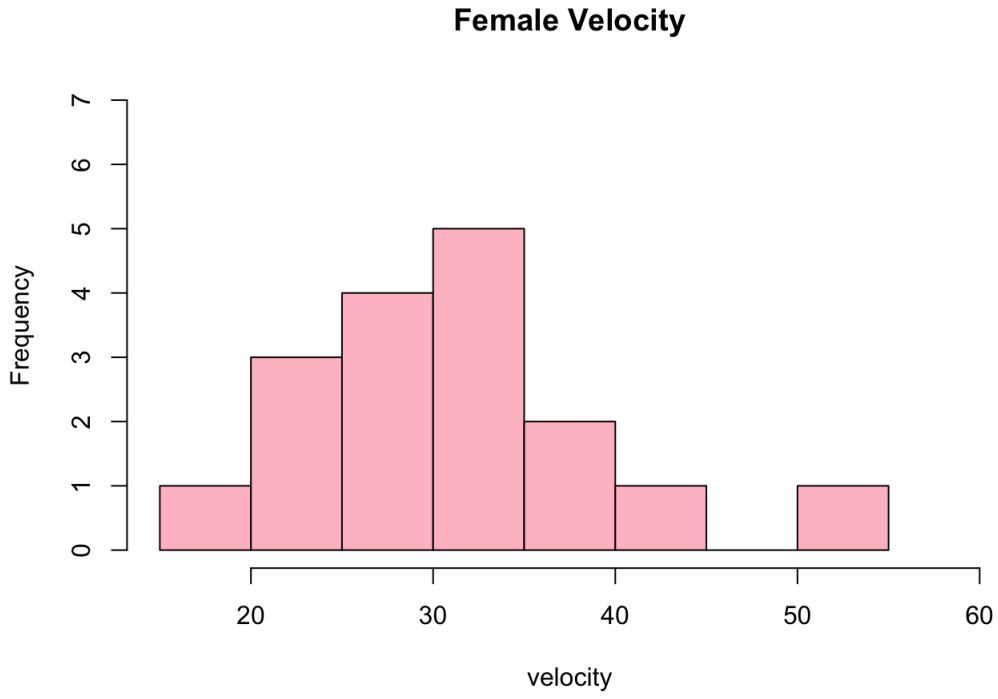
a. Use R to create a histogram for the males and a histogram for the females (any kind of histogram that you want). Adjust the x axis scale so the two groups are more easily compared.

```
?hist
Males = c(54.2, 39.6, 52.3, 48.4, 35.9, 30.4, 25.2, 45.4, 48.9, 48.9, 45.8, 44.0, 52.5, 48.3, 59.9, 51.7, 38
.6, 39.1, 49.9, 38.3)
Females = c(30.3, 43.0, 25.7, 26.7, 27.3, 31.9, 53.7, 32.9, 19.4, 23.7, 23.3, 23.3, 37.8, 39.5, 33.5, 30.4,
28.53)

hist(Males, main = "Male Velocity", xlab = "velocity",col = "blue", xlim = c(15,60), ylim = c(0,7))
```




```
hist(Females, main = "Female Velocity", xlab = "velocity",col = "pink", xlim = c(15,60), ylim = c(0,7))
```



b. Compare the shape of the throws from the male and female students observed in this sample.

Females is right skewed. Males is left skewed.

c. Compute and compare the mean and median throw velocities observed in the male and famale students across gender.

```
mean(Males)
```

```
## [1] 44.865
```

```
median(Males)
```

```
## [1] 47.05
```

```
mean(Females)
```

```
## [1] 31.23118
```

```
median((Females))
```

```
## [1] 30.3
```

The mean and median of the males data is higher.

d. Compute and compare the standard deviation in throw velocities observed in the male and famale students.

```
sd(Males)
```

```
## [1] 8.513845
```

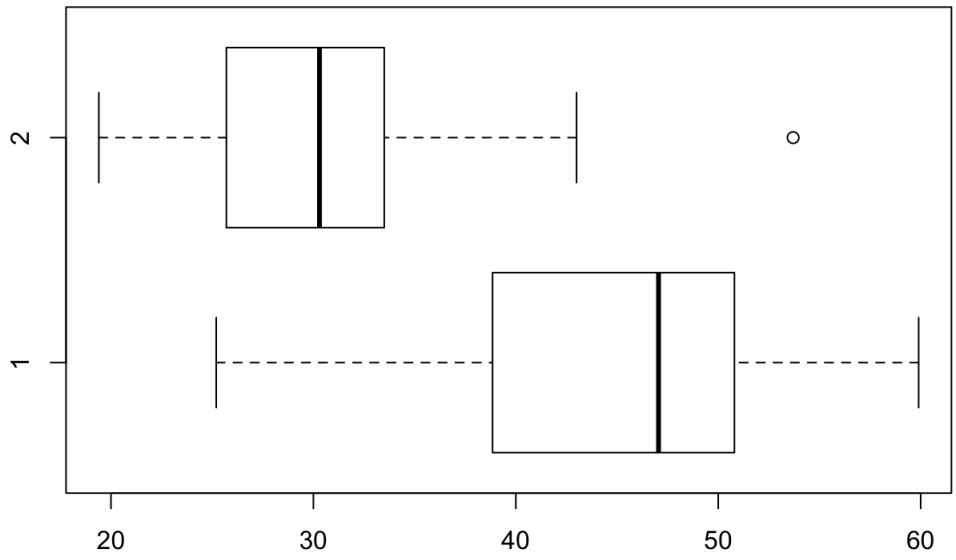
```
sd(Females)
```

```
## [1] 8.519068
```

The standard deviation of both the data sets are about the same.

e. Use R (or by hand if you prefer) to help you create a boxplots of the two sets so they are easily comparable.

```
data =range(Males,Females)
boxplot(Males, Females, horizontal = T)
```



f. Which, if any values were identified as outliers? Would this value have been identified as an outlier if it were thrown by the opposite gender?

Female = 53.7 Male = 25.2

These values are the biggest outliars in the data. If the female and male outliars were to swap, then they would not be outliars.

Problem 6

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?

```
((-11.9 + 1.19) / 12)
```

```
## [1] -0.8925
```

```
24 - 0.8925
```

```
## [1] 23.1075
```

The mean will decrease by 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?

Yes, since the smallest number had been decreased it remained the smallest number. The smallest number in the dataset does not effect the median so the median didn't change.

c. Is it possible to determine the direction in which the standard deviation changes? Or how much the standard deviation changes? If so, by how much does it change? If not, why not?

```
((-11.9^2 + 1.19^2) + (12 * (24^2 - 23.1075^2))) * (1 / (12 - 1))
```

```
## [1] 33.12067
```

No, we cannot determine how much the standard deviation will change. But we can determine the square of the standard deviation will increase by 33.12067.

Problem 7

The UW Statisics Department is trying to determine what day of the week to hold their annual fall festival. Assume that the weekdays, Monday through Friday, are equally likely and that each weekend day, Saturday and Sunday, is three times as likely as a weekday to be selected.

a. Assign probabilities to the seven outcomes.

Monday = 1

Teusday = 1

Wednesday = 1

Thursday = 1

Friday = 1

Saturday = 3

Sunday = 3

b. Find the probability a weekday will be selected.

Weekday = 5 / 11

Problem 8

(*)Suppose you are eating at a pizza parlor with two friends. You have agreed to the following rule to decide who will pay the bill. Each person will toss a coin. The person who gets a result that is different from the other two will pay the bill. If all three tosses yield the same result, the bill will be shared by all. (It may be helpful to list the outcomes in the sample space)

a. Find the probability that only you have to pay.

HHH, HHT, HTH, THH, HTT, THT, TTH, TTT

THH, HTT = I pay is 1/4

b. Find the probability that all three will share.

HHH, TTT = share bill is 1/4

Problem 9

(*) The following frequency table shows the classification of 58 landfills in a state according to their concentration of the three hazardous chemicals arsenic, barium, and mercury. If a landfill is selected at random, find the probability that it has:

a. A high concentration of barium.

s = 5 + 11

m = 58

The probability that its concentration is high in Barium is 16 / 58

```
16 / 58
```

```
## [1] 0.2758621
```

b. A high concentration of mercury and low concentrations of both arsenic and barium.

s = 10

m = 58

The probability that its concentration is high in Mercury and low in Arsenis and barium is 10 / 58

```
10 / 58
```

```
## [1] 0.1724138
```

c. A high concentration of any one of the chemicals and low concentrations of the other two.

s = 9 + 8 + 10

m = 58

The probability that its concentration is high in one chemical and low in the other two is 27 / 58

```
27 / 58
```

```
## [1] 0.4655172
```

d. A high concentration of Mercury given it has a high concentration of Arsenic. Is having high concentrations of Mercury and high concentrations of Arsenic independent in this set of landfills?

P(High Mercury | High Arsenic) = P(High Mercury and Arsenic) / P(High Arsenic) = 0.3333

The probability that its concentration is high in Mercury given that it is a high concentration of Arsenic is 0.3333

```
((1 + 5) / 58) / (18 / 58)
```

```
## [1] 0.3333333
```

e. A high concentration of Barium given it has a low concentration of Mercury. Is having high concentrations of Barium and low concentrations of Mercury independent in this set of landfills?

P(High Barium | Low Mercury) = P(High Barium and Low Mercury) / P(Low Mercury) = 0.2894737

The probability that its concentration is high in Barium given that it is a low concentration of Mercury is 0.2894737

```
((3 + 8) / 58) / ((11 + 27) / 58)
```

```
## [1] 0.2894737
```