Timothy Simmons
ti812979
STA 3032
MW 4:30 – 5:45 PM

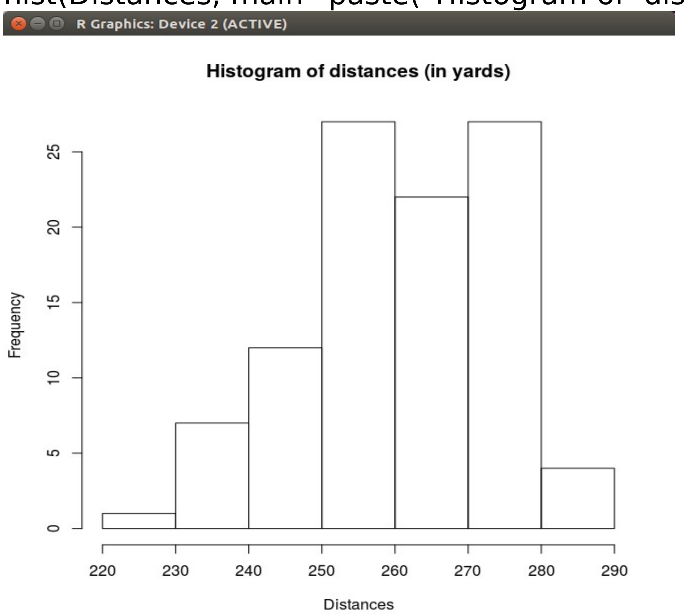
Final Term Project

1.) For easier processing using R, I transferred the data to a file called “golf”.

   a. In R, loaded data using “x <- read.csv(“golf”, header=F, skip=’\n’)
      i. Since read.csv returns a 'List', take the list and store just the vector V1: “Distances <- x$V1”

   b. The sample statistics:
      i. Sample mean
         1. mean(Distances) = 260.302
      ii. Sample standard deviation
         1. sd(Distances) = 13.40828
      iii. Sample median
         1. median(Distances, na.rm=T)
      iv. Third quartile
         1. quantile(Distances)[4]

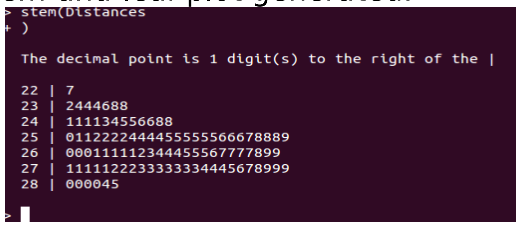
   c. Histogram of overall distance
      i. hist(Distances, main=paste(“Histogram of distances (in yards)”))



      ii.

d. stem(Distances)

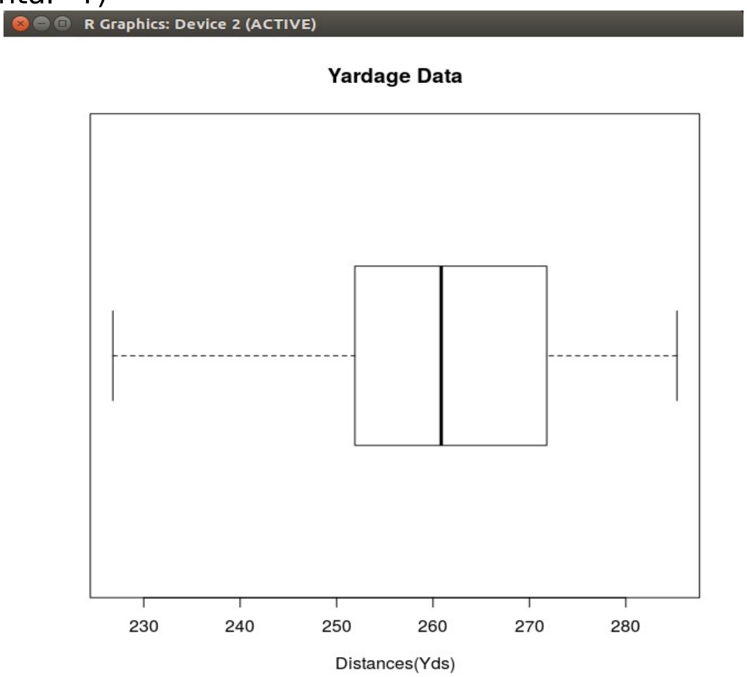
   i. The stem-and-leaf plot generated:



```
> stem(Distances
+ )

  The decimal point is 1 digit(s) to the right of the |

  22 | 7
  23 | 2444688
  24 | 111134556688
  25 | 0112222444455555556667888 9
  26 | 00011111234445556777789 9
  27 | 1111122233333334445678999
  28 | 000045

>
```

The steam and leaf plot seems to show the same trend as the histogram, as it looks like the histogram laid on its right side.

e. Boxplot(Distances, main=”Yardage Data”, xlab=”Distance(Yds)”, horizontal=T)



   i. The boxplot shows the same stuff that the previous plots, and the calculations showed: there seems to be a good bit of variability in the data, with the mean and most of the values centered around 260.

f. $\bar{y} \pm 2s : [233.4854, 287.1186]$

2.) For easier processing in R, stored COGAS table in file called “COGAS”.

a. Read into R using “x ← read.csv(“COGAS”, sep='\t')”

b. While building the table, we're only interested in a specific year

   In order to get only those years, we filter using:

   x ← x[x$V1 == 2001, 6] #or 2000 if we're searching for it

   The first bit is the slice which only includes the proper year

   The second strips out all but the CO-ppb

Statistic Format is Name of statistic: command/formula used	Year	
	2000	2001
Mean: mean(data)	173.0683	187.4107
Median: median(data)	172.79	185.165
1Q: quantile(data)[2]	170.67	181.79
3Q: quantile(data)[4]	174.97	192.22
Range: (min(data), max(data))	(169.04 , 178.59)	(178.59 , 203.29)
IQR: 3Q- 1Q	4.3	10.43
Variance: var(data)	6.930317	46.49976
Standard Deviation: sd(data)	2.632549	6.819073
Standard Error: $\sqrt{\operatorname{var}(data) / \operatorname{len}(data)}$	0.1940741	.4683359

3.) Read in data stored in “Visit”

```
x ← read.table(“Visit”, sep='\t')
```

Filter data into vector for the pie chart

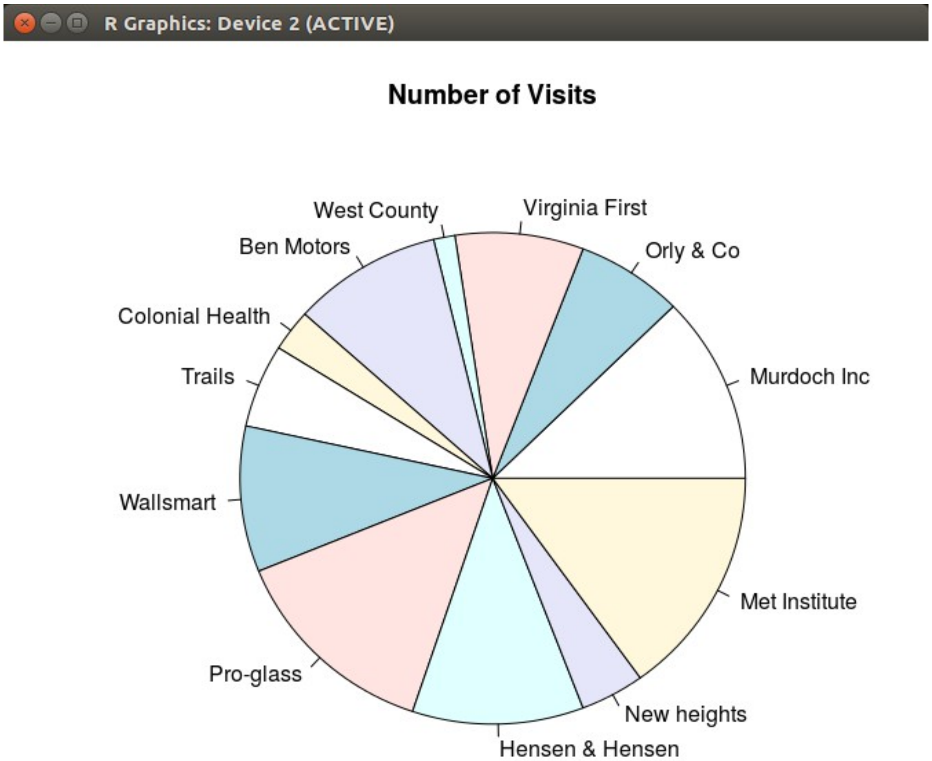
```
data ← as.integer(x[-1,2]) #Skip the header, take the #visits column
```

Filter the names of the companies into our label

```
lbls ← x[-1,1]
```

Display pie chart:

```
pie(data, labels = lbls, main = x[1,2])
```



4.) The appropriate probability distribution for Y is the Poisson distribution, with μ = *s* = 18 pphm. The justification is Y is based off a Bernoulli trial, but we do not have enough information to work with a binomial distribution. There may also be a clue in that the mean is given as "number of exceedances in a year". The sample passes the sample size requirement of the Poisson distribution: n is approximately 365, which is well above 50. The problem is we don't know what the probability is, but as long as the probability is greater than 1.3% it still holds, since 0.013 * 365 = 4.745 and 0.014 * 365 = 5.11

a. $P(Y \leq 20) = \sum_{0}^{20} \mathrm{p}(x;18) = \sum_{0}^{20} \frac{e^{-18} 18^{x}}{x!} = .731$

b. $P(5 \leq Y \leq 20) = \sum_{5}^{20} \frac{e^{-18} 18^{x}}{x!} = .731$

c. For a Poisson distribution, the variance is equal to the mean, so the estimate of the standard deviation would be 4.24, the square root of 18. So, we would expect Y to fall within the range [13.76, 22.24].

5.) Standard deviation: 500 km
G1 Mean : 1500 km
G2 Mean : 1200 km

a. Assuming a mean displacement of 1500 km, there is a 2.28% chance of a displacement of 500 km.

$$Z = \frac{X-1500}{500} = \frac{500-1500}{500} = -2$$

$$P(X<500) = \Phi(-2) = .0228$$

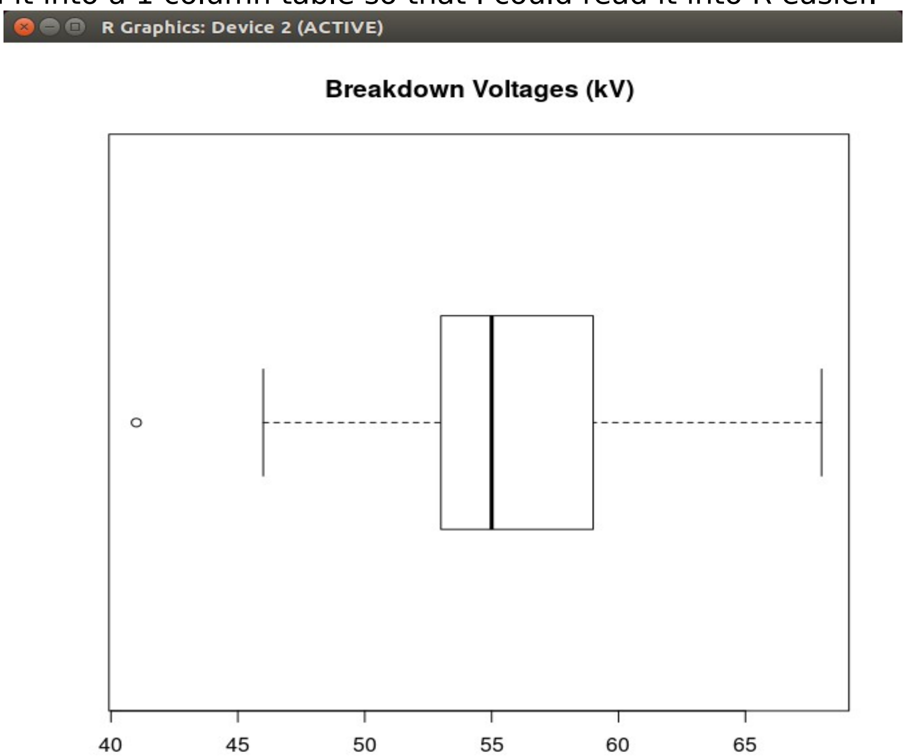
b. Assuming a mean displacement of 1200km, there is a 8.08% chance of a displacement of 500 km.

$$Z = \frac{X-1200}{500} = \frac{500-1200}{500} = -1.4$$

$$P(X<500) = \Phi(-1.4) = .0808$$

c. Based on the above probabilities, the most plausible mean is 1200km.

6.) As usual, transferred the data to a file called “AC” for easier processing in R. Since there are no headers, I assumed the entire set is supposed to be a list, so I turned it into a 1 column table so that I could read it into R easier.



a.

It looks like the minimum value, 41, is an outlier. The data seems to be skewed to the left, with there being a longer right-tail, in both the third quartile and the right whisker. The mean looks to be about 55.

$$Z_{a/2}=1.645$$

b.
$$55-1.645*\frac{5.62}{\sqrt{48}}\leq\mu\leq 55+1.645*\frac{5.62}{\sqrt{48}}$$

$$53.66561\leq\mu\leq 56.33439 \text{ with 90\% confidence}$$

The range of possible values in the estimation is fairly narrow, so it would seem that μ was precisely estimated.

c.
$$1.96*\frac{5.62}{\sqrt{n}}=1$$

$$\frac{\sqrt{n}}{1.96*5.62}=1$$

$$\sqrt{n}=1.96*5.62$$

$$n=(1.96*5.62)^2=121\text{, rounded down}$$

The sample size required to have a 2kV width on the CI is 121.

7.) For simplicity, assume the distribution of the data is approximately normal, X ~ n(2571.429, 13247.62).
H_0: μ = 2500
H_a: μ > 2500
We're testing whether or not the pipes have a mean greater than 2500, so we test for mean = 2500 and determine if there's reason to dismiss the null hypothesis.

$$Z=\frac{X-2500}{115.0983/\sqrt{7}}=1.64193$$

$$1-PHI(Z)=1-.9495=.0505$$

Assuming a 95% Confidence Level, which is probably a fairly good assumption, since it would only pass a lower confidence level, the sample doesn't meet the required specifications. There is no evidence that the sample pipes have a mean breaking strength of more than 2500: since p > α, there is not enough evidence to reject the null hypothesis.