STAT253 â€“ Homework1

Berk YILDIZ - 150117052

23 Mart 2020

# Import data and create vector  
data <- read.csv(file="D:/berky/Documents/R/HW#1/HW1\_Data.csv",header=TRUE,sep=";")  
# 1.Variables are quantiative discrete.Data are multivariate.  
# 2.Pie chart, bar chart,dotplots, stem and leaf plots can be used. We can see how are data's location, spread, shape and are there any outliers.  
# 3. a.) Calculate the sample mean.  
tapply(data$SYSBP, data$GENDER, mean)

## 0 1   
## 117.07 110.48

tapply(data$DIASBP, data$GENDER, mean)

## 0 1   
## 74.25 71.75

# b.) Calculate the sample variance.  
tapply(data$SYSBP, data$GENDER, var)

## 0 1   
## 182.7728 183.7269

tapply(data$DIASBP, data$GENDER, var)

## 0 1   
## 107.17929 96.79545

# c.) Find the sample standard deviation.  
sdsysbp <- tapply(data$SYSBP, data$GENDER, sd)  
sdsysbp

## 0 1   
## 13.51935 13.55459

sddiasbp <- tapply(data$DIASBP, data$GENDER, sd)  
sddiasbp

## 0 1   
## 10.352743 9.838468

# d.) Find the lower and upper quartiles.  
# Lower quartile, is the value that cuts off the first 25% of the data when it is sorted in ascending order.Upper quartile, is the value that cuts off the first 75%.  
tapply(data$SYSBP, data$GENDER, quantile)

## $`0`  
## 0% 25% 50% 75% 100%   
## 70 110 118 124 170   
##   
## $`1`  
## 0% 25% 50% 75% 100%   
## 80 100 110 120 150

tapply(data$DIASBP, data$GENDER, quantile)

## $`0`  
## 0% 25% 50% 75% 100%   
## 46 70 74 80 100   
##   
## $`1`  
## 0% 25% 50% 75% 100%   
## 48.0 67.5 70.0 76.0 108.0

# e.) What are the maximum and minimum values?  
minmaxsysbp <- tapply(data$SYSBP, data$GENDER, range)  
minmaxsysbp

## $`0`  
## [1] 70 170  
##   
## $`1`  
## [1] 80 150

minmaxdiasbp <- tapply(data$DIASBP, data$GENDER, range)  
minmaxdiasbp

## $`0`  
## [1] 46 100  
##   
## $`1`  
## [1] 48 108

# f.) Calculate the range.  
rangesysbp <- lapply(minmaxsysbp,diff)  
rangesysbp

## $`0`  
## [1] 100  
##   
## $`1`  
## [1] 70

rangediasbp <- lapply(minmaxdiasbp,diff)  
rangediasbp

## $`0`  
## [1] 54  
##   
## $`1`  
## [1] 60

# g.) Compare the range and the standard deviation. The range is approximately how many standard deviations?  
100/13.51935

## [1] 7.396805

70/13.55459

## [1] 5.164302

54/10.352743

## [1] 5.216009

60/9.838468

## [1] 6.09851

# h.) What is the median?  
tapply(data$SYSBP, data$GENDER, median)

## 0 1   
## 118 110

tapply(data$DIASBP, data$GENDER, median)

## 0 1   
## 74 70

# i) Calculate the IQR.  
tapply(data$SYSBP, data$GENDER, IQR, na.rm = FALSE)

## 0 1   
## 14 20

tapply(data$DIASBP, data$GENDER, IQR, na.rm = FALSE)

## 0 1   
## 10.0 8.5

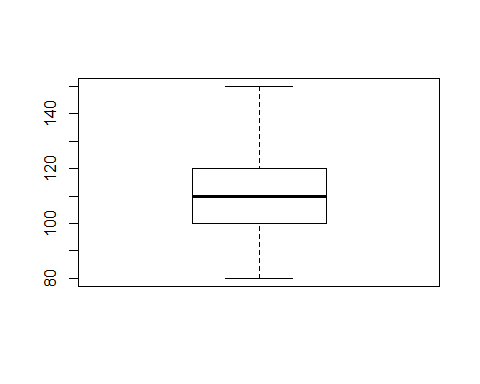
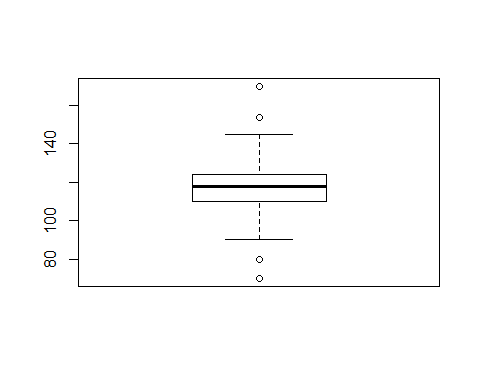
# j.) Calculate five-number summaries.  
tapply(data$SYSBP, data$GENDER, fivenum)

## $`0`  
## [1] 70 110 118 124 170  
##   
## $`1`  
## [1] 80 100 110 120 150

tapply(data$DIASBP, data$GENDER, fivenum)

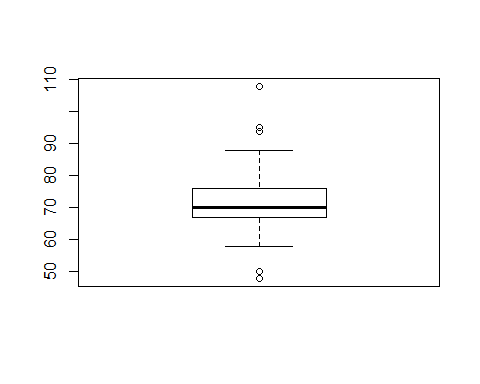
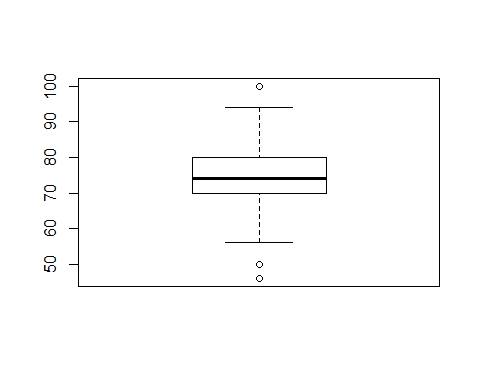
## $`0`  
## [1] 46 70 74 80 100  
##   
## $`1`  
## [1] 48 67 70 76 108

# k.) Construct box plot for the data set. Are there any outliers? What does the box plot tell you about the shape of the distribution?  
tapply(data$SYSBP, data$GENDER, boxplot)



## $`0`  
## $`0`$stats  
## [,1]  
## [1,] 90  
## [2,] 110  
## [3,] 118  
## [4,] 124  
## [5,] 145  
## attr(,"class")  
## 1   
## "integer"   
##   
## $`0`$n  
## [1] 100  
##   
## $`0`$conf  
## [,1]  
## [1,] 115.788  
## [2,] 120.212  
##   
## $`0`$out  
## [1] 70 154 80 170  
##   
## $`0`$group  
## [1] 1 1 1 1  
##   
## $`0`$names  
## [1] "1"  
##   
##   
## $`1`  
## $`1`$stats  
## [,1]  
## [1,] 80  
## [2,] 100  
## [3,] 110  
## [4,] 120  
## [5,] 150  
## attr(,"class")  
## 1   
## "integer"   
##   
## $`1`$n  
## [1] 100  
##   
## $`1`$conf  
## [,1]  
## [1,] 106.84  
## [2,] 113.16  
##   
## $`1`$out  
## numeric(0)  
##   
## $`1`$group  
## numeric(0)  
##   
## $`1`$names  
## [1] "1"

tapply(data$DIASBP, data$GENDER, boxplot)



## $`0`  
## $`0`$stats  
## [,1]  
## [1,] 56  
## [2,] 70  
## [3,] 74  
## [4,] 80  
## [5,] 94  
## attr(,"class")  
## 1   
## "integer"   
##   
## $`0`$n  
## [1] 100  
##   
## $`0`$conf  
## [,1]  
## [1,] 72.42  
## [2,] 75.58  
##   
## $`0`$out  
## [1] 50 50 46 100  
##   
## $`0`$group  
## [1] 1 1 1 1  
##   
## $`0`$names  
## [1] "1"  
##   
##   
## $`1`  
## $`1`$stats  
## [,1]  
## [1,] 58  
## [2,] 67  
## [3,] 70  
## [4,] 76  
## [5,] 88  
## attr(,"class")  
## 1   
## "integer"   
##   
## $`1`$n  
## [1] 100  
##   
## $`1`$conf  
## [,1]  
## [1,] 68.578  
## [2,] 71.422  
##   
## $`1`$out  
## [1] 95 94 48 50 94 95 50 108  
##   
## $`1`$group  
## [1] 1 1 1 1 1 1 1 1  
##   
## $`1`$names  
## [1] "1"

# k1. There are 4 outliers. Shape of the distribution is skewed left.  
# k2. There aren't any outliers. Shape of the distribution is symmetric.  
# k3. There are 3 outliers. Shape of the distribution is skewed right.  
# k3. There are 5 outliers. Shape of the distribution is skewed right.  
# l.) Construct and interpret stem-and-leaf plot.  
tapply(data$SYSBP, data$GENDER, stem)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 6 | 0  
## 8 | 006  
## 10 | 00000022466688888000000000000002244444455668888888  
## 12 | 0000000000000022244444445688888000002226788  
## 14 | 54  
## 16 | 0  
##   
##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 8 | 008  
## 9 | 00004456888888  
## 10 | 0000000000000222224466688  
## 11 | 000000000000000000022245556668  
## 12 | 000000000004466688  
## 13 | 00006  
## 14 | 0002  
## 15 | 0

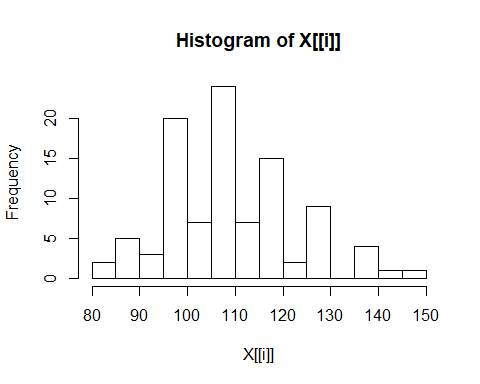
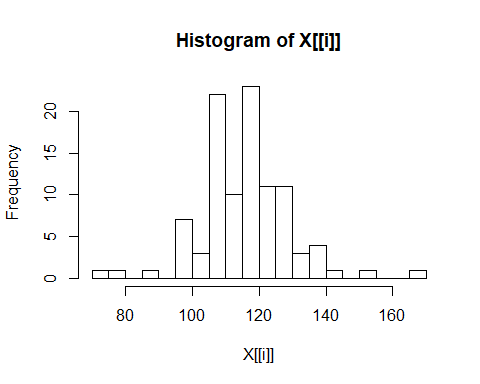
## $`0`  
## NULL  
##   
## $`1`  
## NULL

tapply(data$DIASBP, data$GENDER, stem)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 4 | 6  
## 5 | 00666  
## 6 | 00000000666668888  
## 7 | 000000000000000000002222244444445668888888  
## 8 | 000000000000224444444468  
## 9 | 0000000224  
## 10 | 0  
##   
##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 4 | 8  
## 5 | 0088  
## 6 | 00000000024444444566888888888  
## 7 | 000000000000000000000000000344444566666666888  
## 8 | 0000000000246888  
## 9 | 4455  
## 10 | 8

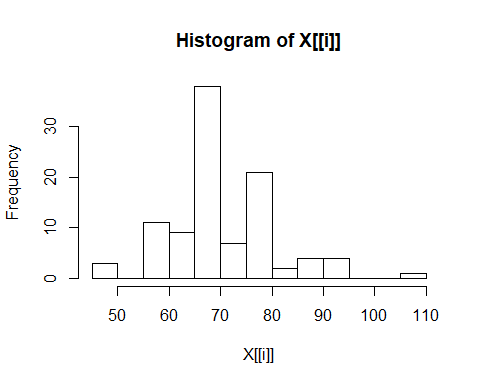
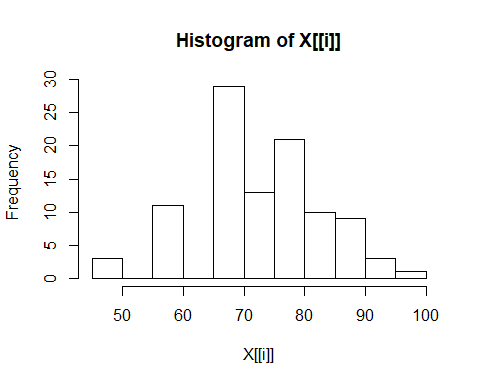
## $`0`  
## NULL  
##   
## $`1`  
## NULL

# l1. There are outliers. Shape of the distribution is skewed left.  
# l2. There aren't any outliers. Shape of the distribution is symmetric.  
# l3. There are outliers. Shape of the distribution is skewed right.  
# l3. There are outliers. Shape of the distribution is skewed right.  
# m.) Construct and interpret histograms. In this solution, you will use different number of subintervals to construct the histograms and then compare the results  
# m1.) use 5 subintervals to construct the histograms and then interpret the graph  
tapply(data$SYSBP, data$GENDER, hist,breaks=22)



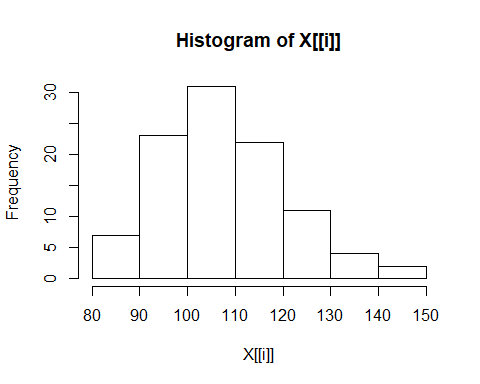
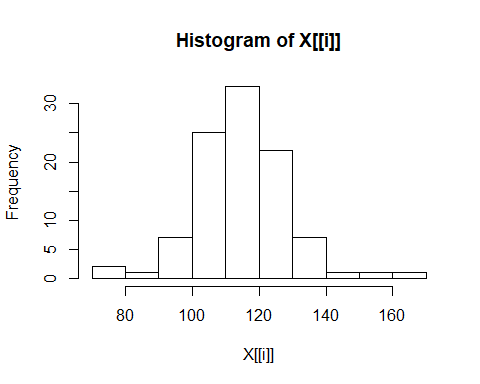
## $`0`  
## $breaks  
## [1] 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150  
## [18] 155 160 165 170  
##   
## $counts  
## [1] 1 1 0 1 0 7 3 22 10 23 11 11 3 4 1 0 1 0 0 1  
##   
## $density  
## [1] 0.002 0.002 0.000 0.002 0.000 0.014 0.006 0.044 0.020 0.046 0.022  
## [12] 0.022 0.006 0.008 0.002 0.000 0.002 0.000 0.000 0.002  
##   
## $mids  
## [1] 72.5 77.5 82.5 87.5 92.5 97.5 102.5 107.5 112.5 117.5 122.5  
## [12] 127.5 132.5 137.5 142.5 147.5 152.5 157.5 162.5 167.5  
##   
## $xname  
## [1] "X[[i]]"  
##   
## $equidist  
## [1] TRUE  
##   
## attr(,"class")  
## [1] "histogram"  
##   
## $`1`  
## $breaks  
## [1] 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150  
##   
## $counts  
## [1] 2 5 3 20 7 24 7 15 2 9 0 4 1 1  
##   
## $density  
## [1] 0.004 0.010 0.006 0.040 0.014 0.048 0.014 0.030 0.004 0.018 0.000  
## [12] 0.008 0.002 0.002  
##   
## $mids  
## [1] 82.5 87.5 92.5 97.5 102.5 107.5 112.5 117.5 122.5 127.5 132.5  
## [12] 137.5 142.5 147.5  
##   
## $xname  
## [1] "X[[i]]"  
##   
## $equidist  
## [1] TRUE  
##   
## attr(,"class")  
## [1] "histogram"

tapply(data$DIASBP, data$GENDER, hist,breaks=10)



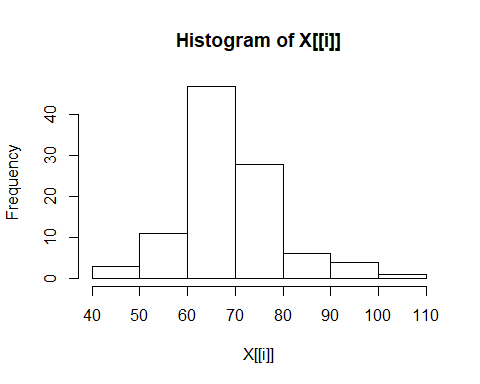
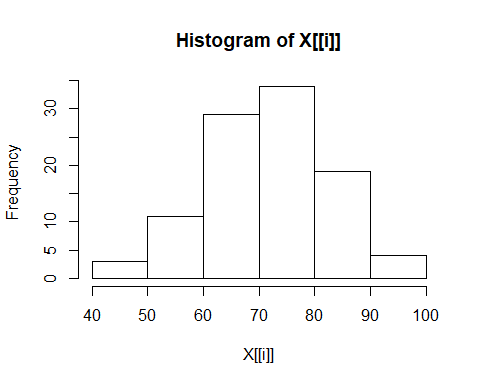
## $`0`  
## $breaks  
## [1] 45 50 55 60 65 70 75 80 85 90 95 100  
##   
## $counts  
## [1] 3 0 11 0 29 13 21 10 9 3 1  
##   
## $density  
## [1] 0.006 0.000 0.022 0.000 0.058 0.026 0.042 0.020 0.018 0.006 0.002  
##   
## $mids  
## [1] 47.5 52.5 57.5 62.5 67.5 72.5 77.5 82.5 87.5 92.5 97.5  
##   
## $xname  
## [1] "X[[i]]"  
##   
## $equidist  
## [1] TRUE  
##   
## attr(,"class")  
## [1] "histogram"  
##   
## $`1`  
## $breaks  
## [1] 45 50 55 60 65 70 75 80 85 90 95 100 105 110  
##   
## $counts  
## [1] 3 0 11 9 38 7 21 2 4 4 0 0 1  
##   
## $density  
## [1] 0.006 0.000 0.022 0.018 0.076 0.014 0.042 0.004 0.008 0.008 0.000  
## [12] 0.000 0.002  
##   
## $mids  
## [1] 47.5 52.5 57.5 62.5 67.5 72.5 77.5 82.5 87.5 92.5 97.5  
## [12] 102.5 107.5  
##   
## $xname  
## [1] "X[[i]]"  
##   
## $equidist  
## [1] TRUE  
##   
## attr(,"class")  
## [1] "histogram"

# m2.) use more subintervals (more than 5 subintervals) to construct the graph and then interpret the graph  
tapply(data$SYSBP, data$GENDER, hist)



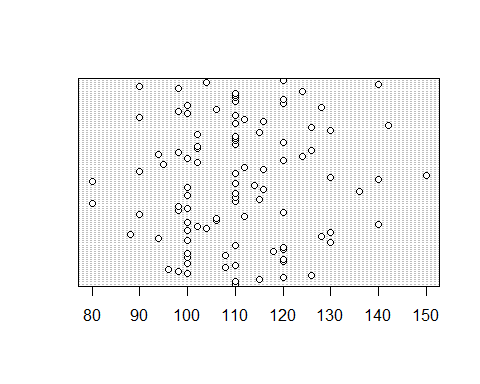
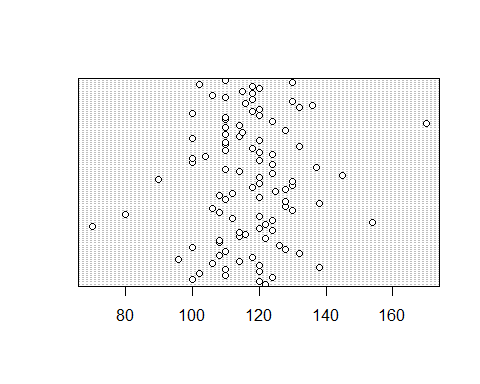
## $`0`  
## $breaks  
## [1] 70 80 90 100 110 120 130 140 150 160 170  
##   
## $counts  
## [1] 2 1 7 25 33 22 7 1 1 1  
##   
## $density  
## [1] 0.002 0.001 0.007 0.025 0.033 0.022 0.007 0.001 0.001 0.001  
##   
## $mids  
## [1] 75 85 95 105 115 125 135 145 155 165  
##   
## $xname  
## [1] "X[[i]]"  
##   
## $equidist  
## [1] TRUE  
##   
## attr(,"class")  
## [1] "histogram"  
##   
## $`1`  
## $breaks  
## [1] 80 90 100 110 120 130 140 150  
##   
## $counts  
## [1] 7 23 31 22 11 4 2  
##   
## $density  
## [1] 0.007 0.023 0.031 0.022 0.011 0.004 0.002  
##   
## $mids  
## [1] 85 95 105 115 125 135 145  
##   
## $xname  
## [1] "X[[i]]"  
##   
## $equidist  
## [1] TRUE  
##   
## attr(,"class")  
## [1] "histogram"

tapply(data$DIASBP, data$GENDER, hist,breaks = 5)



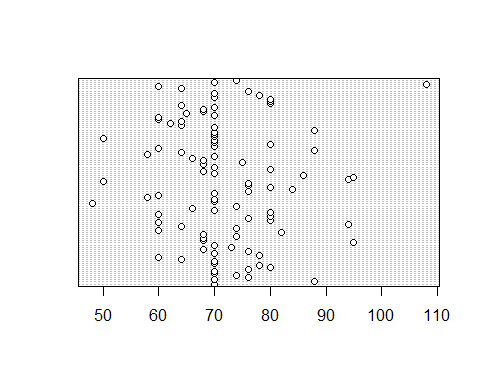
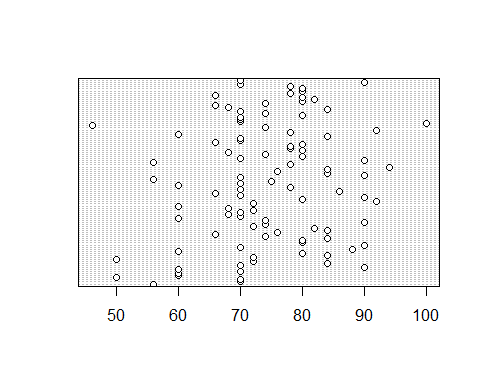
## $`0`  
## $breaks  
## [1] 40 50 60 70 80 90 100  
##   
## $counts  
## [1] 3 11 29 34 19 4  
##   
## $density  
## [1] 0.003 0.011 0.029 0.034 0.019 0.004  
##   
## $mids  
## [1] 45 55 65 75 85 95  
##   
## $xname  
## [1] "X[[i]]"  
##   
## $equidist  
## [1] TRUE  
##   
## attr(,"class")  
## [1] "histogram"  
##   
## $`1`  
## $breaks  
## [1] 40 50 60 70 80 90 100 110  
##   
## $counts  
## [1] 3 11 47 28 6 4 1  
##   
## $density  
## [1] 0.003 0.011 0.047 0.028 0.006 0.004 0.001  
##   
## $mids  
## [1] 45 55 65 75 85 95 105  
##   
## $xname  
## [1] "X[[i]]"  
##   
## $equidist  
## [1] TRUE  
##   
## attr(,"class")  
## [1] "histogram"

# m3.) compare two results above and discuss which one is better and why.  
#Subintervals with 5 is more detailed than subintervals with 10. Lower subintervals are better.  
# n.) Draw a dotplot of this data set. Are the data mound shaped?  
#Only women systolic blood pressure data set is mound shaped.  
# o.) Can you use Tchebysheff’s Theorem to describe this data set? Why or why not?  
#Yes, Tchebysheff’s Theorem must be true for any data set.  
# p.) Can you use the Empirical Rule to describe this data set? Why or why not?  
#Yes just for the second one.Only women systolic blood pressure data set is mound shaped.  
# q.) Given data sets of two groups, construct and interpret a comparative dotplot to compare the groups  
tapply(data$SYSBP, data$GENDER, dotchart)



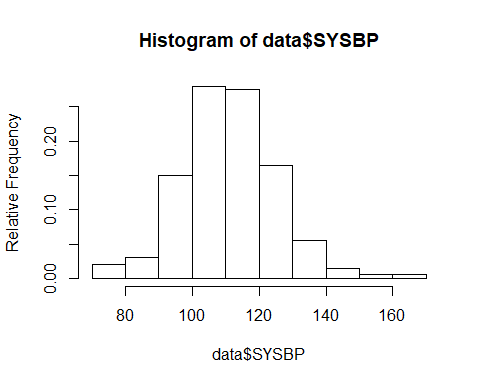
## $`0`  
## NULL  
##   
## $`1`  
## NULL

tapply(data$DIASBP, data$GENDER, dotchart)

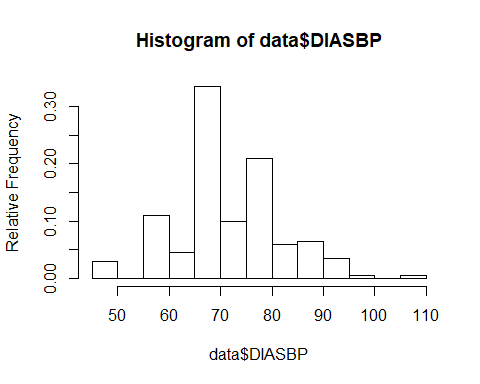


## $`0`  
## NULL  
##   
## $`1`  
## NULL

# Second one is more widely distrubuted than first one.Third one is more in centered than fourth.  
# r.) Construct a relative frequency histogram for these data. How would you describe the shape of the distribution?  
h<-hist(data$SYSBP, plot=F)  
h$counts <- h$counts / sum(h$counts)  
plot(h, freq=TRUE, ylab="Relative Frequency")



#Symmetric distrubution  
h<-hist(data$DIASBP, plot=F)  
h$counts <- h$counts / sum(h$counts)  
plot(h, freq=TRUE, ylab="Relative Frequency")



#Skewed right  
# s.) Find the z-scores for the largest and smallest observations. Would you consider them to be outliers? Why or why not?  
(70-117.07)/13.51935

## [1] -3.481676

(170-117.07)/13.51935

## [1] 3.915129

(80-110.48)/13.55459

## [1] -2.248685

(150-110.48)/13.55459

## [1] 2.915618

(46-74.25)/10.352743

## [1] -2.728745

(100-74.25)/10.352743

## [1] 2.487264

(48-71.75)/9.838468

## [1] -2.413994

(108-71.75)/9.838468

## [1] 3.684517

#First two and the last one is outlier. Because they provide z<-3 or z>3