**Homework #1**

1. Targeted population: Population of USA  
   Sampling population: 34,000 random chosen individuals  
   Sample: Individuals who answered the questionnaire  
   Parameter: Proportion of all smokers who died due to heart disease  
   Statistic: Proportion of smokers who have died due to heart disease in sample  
     
   This is an observational study and not controlled experiment because we are not assigning people to groups.  
     
   We cannot draw a conclusion that smoking causes heart disease as this is not a controlled experiment.
2. Targeted population: Households in the continental United States  
   Sampling population: 500 households in the continental United States  
   Sample: Individuals who answered the 50 questionnaires  
   Parameter: Consumer sentiment score  
   Statistic: Mean of the consumer sentiment score
3. It is given that Mr. Ironheart is being paid 80 million dollar a year. Such a heavy increase can be treated as an outlier and can contribute to a significant increase in the mean. This will not affect the median. Hence, Mr. Ironheart must be using Mean to show average payroll increase. On the other hand, employee union’s president’s metric shows only a 1 % increase in the average employee salary. This clearly means he is talking median as a heavy outlier will significantly increase the mean but not the median. In this case, I believe mean is the right measure of central tendency because addition of employees apart from CEO who have higher salaries cannot significantly increase the median but will definitely affect the median. Hence, Mr. Ironheart’s claim makes more sense even though it is a bit heart breaking to hear ;)
4. The claim looks incorrect for the following reasons:
   1. Class of 1977 of Harvard may not be a good representative of all college students
   2. Total number of dropouts is very much less than that of graduates. So, the data is imbalanced. We must use normalization to compare things here.

The reason why average wealth of dropouts is higher than that of graduates is more because they might have figured out their vision and pursued the entrepreneurial path while in college and so gave up on education. Mark Zuckerberg and Bill Gates are few relevant examples.

MyTime <- scan(file = "CollectedResponseTimes.txt")  
time.data <- data.frame(MyTime)  
time.data$obs.order <- seq(length(time.data$MyTime))  
time.data$obs.phrase <- ifelse(time.data$obs.order <= 10,  
 1,  
 ifelse(time.data$obs.order <= 20,  
 2, 3))  
# scatter plot of ResponseTime vs Observation Order  
plot(MyTime ~ obs.order, data = time.data)

Chart, scatter chart

Description automatically generated

# Stratified scatter plot of 3 stages  
plot(MyTime ~ obs.phrase, data = time.data)

Chart, scatter chart

Description automatically generated

summary(MyTime)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.4640 0.7060 0.8000 0.8744 1.0470 1.4790

summary(MyTime[1:10])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.6560 0.7520 0.7965 0.8479 0.9677 1.1040

summary(MyTime[11:20])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.6250 0.7300 1.0365 0.9809 1.0675 1.4790

summary(MyTime[21:30])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.4640 0.6843 0.7195 0.7944 0.8940 1.2640

The difference in descriptive statistics for all 3 phrases indicate that the response time are not coming from a stationary process.

salary.data <- read.table(file="salary.txt", header=TRUE)  
salary.data

## GENDER SALARY  
## 1 F 100  
## 2 F 95  
## 3 F 105  
## 4 F 105  
## 5 F 110  
## 6 F 98  
## 7 F 105  
## 8 F 125  
## 9 F 130  
## 10 F 200  
## 11 F 120  
## 12 F 115  
## 13 F 110  
## 14 F 130  
## 15 F 120  
## 16 F 115  
## 17 F 110  
## 18 F 120  
## 19 F 115  
## 20 F 150  
## 21 M 150  
## 22 M 205  
## 23 M 210  
## 24 M 220  
## 25 M 205  
## 26 M 225  
## 27 M 230  
## 28 M 240  
## 29 M 220  
## 30 M 230  
## 31 M 235  
## 32 M 225  
## 33 M 230  
## 34 M 250  
## 35 M 245  
## 36 M 230  
## 37 M 225  
## 38 M 220  
## 39 M 180  
## 40 M 221

# a) Histogram of the salaries using R default setting  
hist(salary.data$SALARY)

Chart, histogram

Description automatically generated

# b) Histogram with break points (at least 15 intervals)  
hist(salary.data$SALARY,  
 breaks=95+(0:14)\*12,  
 main = "Histogram Plot - Salaries",  
 xlab = "Salaries",  
 ylab = "Frequency",  
 border = FALSE,  
 labels = TRUE,  
 xlim = c(min(salary.data$SALARY), max(salary.data$SALARY)),  
 ylim = c(0, 8),  
 col = rainbow(12))

Chart, histogram

Description automatically generated

# c) Boxplot of the salaries  
boxplot(salary.data$SALARY,  
 main = "Box Plot - Salaries",  
 xlab = "All genders",  
 ylab = "Salaries",  
 labels = TRUE,  
 boxwex = 0.3,  
 outline = TRUE,  
 las = 1,  
 notch = FALSE,  
 staplewex = 1,  
 col = "purple")

Chart, box and whisker chart

Description automatically generated

# d) Boxplots of the salaries in two gender groups  
boxplot(SALARY~GENDER, data=salary.data,  
 main = "Box Plot - Salaries grouped by Gender",  
 xlab = "Gender",  
 ylab = "Salaries",  
 labels = TRUE,  
 boxwex = 0.3,  
 outline = TRUE,  
 outpch = 16,  
 outcol = "seagreen3",  
 las = 1,  
 notch = FALSE,  
 staplewex = 1,  
 col = "tomato")  
  
library(psych)

Chart, box and whisker chart

Description automatically generated

# e) summary statistics of salaries as one group and summary statistics within each gender  
describe(salary.data$SALARY)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 40 169.35 55.83 165 168.94 81.54 95 250 155 0.02 -1.83 8.83

describeBy(salary.data$SALARY, salary.data$GENDER)

##   
## Descriptive statistics by group   
## group: F  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 20 118.9 23.01 115 114.69 14.83 95 200 105 2.16 5.03 5.15  
## ------------------------------------------------------------   
## group: M  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 20 219.8 22.57 225 223.19 7.41 150 250 100 -1.53 2.37 5.05

1. Looks like b) is a better representation of the histogram plot. Because we are getting even spread of data when the intervals are close to each other. Higher the number of intervals we get a better distribution of the data but with a greater complexity. Hence, the number of intervals has the chosen wisely. In this case, 15 intervals look like a better choice.  
     
   From the box plot grouped by gender, the weekly salaries for the women appear to be much lesser compared to that of men.
2. The central tendencies for example mean, median produced by describe and describe by groups appear to be far apart from each other. This clearly implies that we cannot summarize the center of the distribution of the salaries.
3. From the above box plots grouped by gender, we could see that women have two outliers located above the maximum whisker which is Q3 + 1.5 times the inter quartile range. Their values are 150, 200. Similarly, men have two outliers located below the minimum whisker which is Q1 - 1.5 times the inter quartile range. Their values are 180, 150.

unicef.data <- read.table(file="unicef.txt", na.strings = ".", header=TRUE)  
unicef.data

## nation lowbwt life60 life92  
## 1 Afghanistan 20 33 43  
## 2 Albania 7 62 73  
## 3 Algeria 9 47 66  
## 4 Angola 19 33 46  
## 5 Argentina 8 65 71  
## 6 Armenia NA NA 72  
## 7 Australia 6 71 77  
## 8 Austria 6 69 76  
## 9 Azerbaijan NA NA 71  
## 10 Bangladesh 50 40 53  
## 11 Belarus NA NA 71  
## 12 Belgium 6 70 76  
## 13 Benin NA 35 46  
## 14 Bhutan NA 37 48  
## 15 Bolivia 12 43 61  
## 16 Botswana 8 46 61  
## 17 Brazil 11 55 66  
## 18 Bulgaria 6 68 72  
## 19 Burkina Faso 21 36 48  
## 20 Burundi NA 41 48  
## 21 Cambodia NA 42 51  
## 22 Cameroon 13 39 56  
## 23 Canada 6 71 77  
## 24 Central African Rep. 15 39 47  
## 25 Chad NA 35 47  
## 26 Chile 7 57 72  
## 27 China 9 47 71  
## 28 Colombia 10 57 69  
## 29 Congo 16 42 52  
## 30 Costa Rica 6 62 76  
## 31 Cote d'Ivoire 14 39 52  
## 32 Cuba 8 64 76  
## 33 Czech Rep. NA NA 72  
## 34 Denmark 6 72 76  
## 35 Dominican Rep. 16 52 67  
## 36 Ecuador 11 53 66  
## 37 Egypt 10 46 61  
## 38 El Salvador 11 50 66  
## 39 Eritrea NA NA 47  
## 40 Estonia NA 69 71  
## 41 Ethiopia 16 36 47  
## 42 Finland 4 68 76  
## 43 France 5 70 77  
## 44 Gabon NA 41 53  
## 45 Georgia NA NA 73  
## 46 Germany NA 70 76  
## 47 Ghana 17 45 56  
## 48 Greece 6 69 77  
## 49 Guatemala 14 46 64  
## 50 Guinea 21 34 44  
## 51 Guinea-Bissau 20 34 43  
## 52 Haiti 15 42 56  
## 53 Honduras 9 46 66  
## 54 Hong Kong 8 66 78  
## 55 Hungary 9 68 70  
## 56 India 33 44 60  
## 57 Indonesia 14 41 62  
## 58 Iran 9 50 67  
## 59 Iraq 15 48 66  
## 60 Ireland 4 70 75  
## 61 Israel 7 69 76  
## 62 Italy 5 69 77  
## 63 Jamaica 11 63 73  
## 64 Japan 6 68 79  
## 65 Jordan 7 47 68  
## 66 Kazakhstan NA NA 69  
## 67 Kenya 16 45 59  
## 68 Korea, Dem. NA 54 71  
## 69 Korea, Rep. 9 54 71  
## 70 Kuwait 7 60 75  
## 71 Kyrgyzstan NA NA 66  
## 72 Lao PDR 18 40 51  
## 73 Latvia NA 70 71  
## 74 Lebanon 10 60 68  
## 75 Lesotho 11 43 60  
## 76 Liberia NA 41 55  
## 77 Libyan Arab Jama. NA 47 63  
## 78 Lithuania NA 69 73  
## 79 Madagascar 10 41 55  
## 80 Malawi 20 38 44  
## 81 Malaysia 10 54 71  
## 82 Mali 17 35 46  
## 83 Mauritania 11 35 48  
## 84 Mauritius 9 59 70  
## 85 Mexico 12 57 70  
## 86 Moldova NA NA 68  
## 87 Mongolia 10 47 63  
## 88 Morocco 9 47 63  
## 89 Mozambique 20 37 47  
## 90 Myanmar 16 44 57  
## 91 Namibia 12 42 59  
## 92 Nepal NA 38 53  
## 93 Netherlands NA 73 77  
## 94 New Zealand 6 71 76  
## 95 Nicaragua 15 47 66  
## 96 Niger 15 35 46  
## 97 Nigeria 16 40 52  
## 98 Norway 4 73 77  
## 99 Oman 10 40 69  
## 100 Pakistan 25 43 59  
## 101 Panama 10 61 73  
## 102 Papua New Guinea 23 41 56  
## 103 Paraguay 8 64 67  
## 104 Peru 11 48 64  
## 105 Philippines 15 53 65  
## 106 Poland NA 67 72  
## 107 Portugal 5 63 75  
## 108 Romania 7 65 70  
## 109 Russian Fed. NA NA 69  
## 110 Rwanda 17 42 46  
## 111 Saudi Arabia 7 44 69  
## 112 Senegal 11 37 49  
## 113 Sierra Leone 17 32 43  
## 114 Singapore 7 64 74  
## 115 Slovakia NA NA 72  
## 116 Somalia 16 36 47  
## 117 South Africa NA 49 63  
## 118 Spain 4 69 77  
## 119 Sri Lanka 25 62 71  
## 120 Sudan 15 39 52  
## 121 Sweden 5 73 78  
## 122 Switzerland 5 71 78  
## 123 Syrian Arab Rep. 11 50 67  
## 124 Tanzania 14 41 51  
## 125 Thailand 13 52 69  
## 126 Togo 20 39 55  
## 127 Trinidad and Tobago 10 63 71  
## 128 Tunisia 8 48 68  
## 129 Turkey 8 50 67  
## 130 Turkmenistan NA NA 66  
## 131 USA 7 70 76  
## 132 Uganda NA 43 42  
## 133 Ukraine NA NA 70  
## 134 United Arab Emirates 7 53 71  
## 135 United Kingdom 7 71 76  
## 136 Uruguay 8 68 72  
## 137 Uzbekistan NA NA 69  
## 138 Venezuela 9 60 70  
## 139 Viet Nam 17 44 64  
## 140 Yemen 19 36 52  
## 141 Yugoslavia (former) NA 63 72  
## 142 Zaire 15 41 52  
## 143 Zambia 13 42 45  
## 144 Zimbabwe 14 45 56

# (a) Histogram and boxplot of low birth weight  
hist(unicef.data$lowbwt,  
 breaks=4+(0:14)\*4,  
 main = "Histogram Plot - Low birth weights",  
 xlab = "Low birth weight",  
 ylab = "Frequency",  
 border = FALSE,  
 labels = TRUE,  
 xlim = c(0, 60),  
 ylim = c(0, 40),  
 col = rainbow(7))

Chart, histogram

Description automatically generated

boxplot(unicef.data$lowbwt,  
 main = "Box Plot - Low birth weights",  
 xlab = "All nations",  
 ylab = "Low birth weights",  
 labels = TRUE,  
 boxwex = 0.3,  
 outline = TRUE,  
 outpch = 16,  
 outcol = "seagreen3",  
 las = 1,  
 notch = FALSE,  
 staplewex = 1,  
 col = "tomato")

Chart, box and whisker chart

Description automatically generated

1. The histogram plot clearly shows that the data is right skewed as the tail is elongated toward the right.
2. The box plot shows that two values lie above maximum whisker, which is Q3 + 1.5 times interquartile range. The values are 50, 33.