STA258-Assignment

Maxim	Piatine
Student ID	1005303100

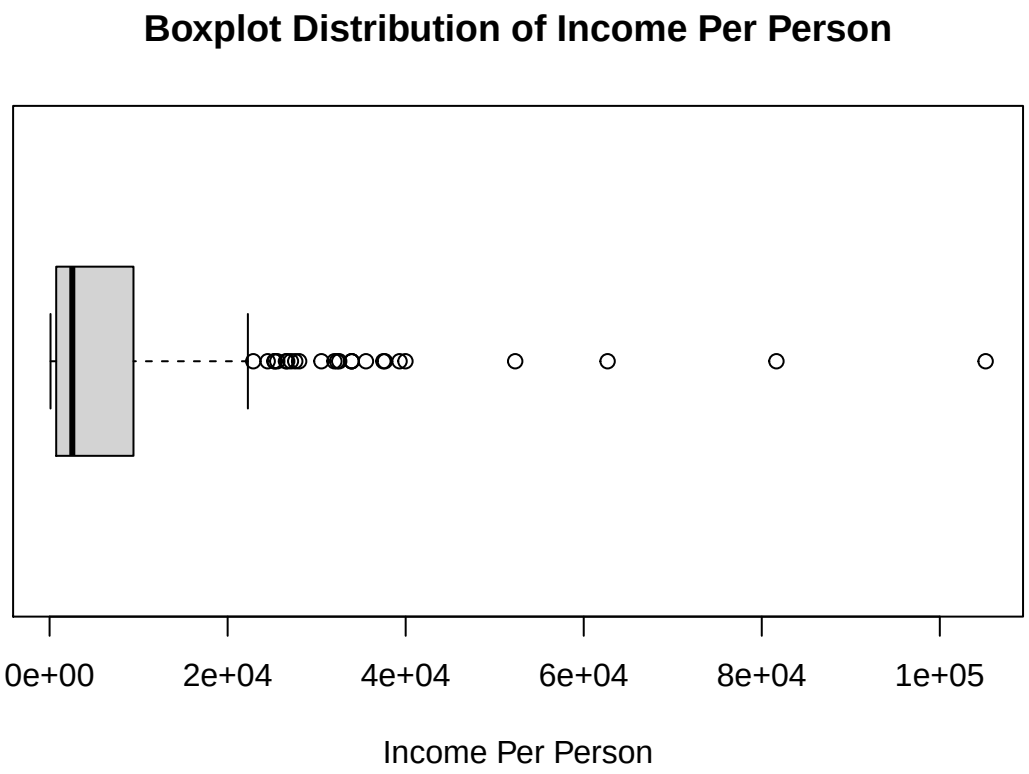
1. (30 points) You have been given gapminder data set. Using R, Evaluate the five number summary for incomeperperson. Graph boxplot for incomeperperson. Looking at the boxplot, do you think mean will be greater than median? Why ? Use no more than two sentences.

   The fivenum function provides an insight on the min, max, and the quartiles. After computing the function, the first number is the minimum, second number is the 1st quartile, the third number is the 2nd quartile/the median, the fourth number is the 3rd quartile, and the fifth number is the maximum.

```
mydata<-read.csv("gapminder.csv")
fivenum(mydata$incomeperperson,na.rm = TRUE)
```

```
## [1]    103.7759    744.2394   2553.4961   9425.3259 105147.4377
```

```
boxplot(mydata$incomeperperson, na.rm=TRUE, horizontal= TRUE,
        xlab='Income Per Person',
        main='Boxplot Distribution of Income Per Person')
```



Based on the box plot, the mean is greater than the median because the distribution is positively skewed because of the median being closer to the first/bottom quartile. Moreover, the data points are mostly clustered around the left tail; however, the right tail of the distribution is much larger.

2. (15 points) Suppose that number of days it takes to recover from illness caused due to coronavirus is normally distributed with mean 14 days and standard deviation of 2 days.

(1) What is the probability that it will take more than 15 days to recover?

```
1-pnorm(15, mean=14, sd=2)
```

```
## [1] 0.3085375
```

(2) What is the probability that it will take less than 11 days to recover?

```
pnorm(11, mean=14, sd=2)
```

```
## [1] 0.0668072
```

(3) What is the probability that it takes between 12 days and 15 days to recover?

```
pnorm(15, mean=14, sd=2)-pnorm(12, mean=14, sd=2)
```

```
## [1] 0.5328072
```

3. (35 points) Data was collected from four STA258 students. The number of text messages these students send while attending 10 STA258 classes is given in the following table:

Winston	Marija	Raiyan	Valerie
65	63	60	57
59	65	56	56
57	59	59	59
58	64	53	56
56	62	56	58
59	68	52	60
62	63	56	59
57	66	58	56
57	65	62	62
56	72	50	61

Is there any evidence that average number of text messages of at least one of the students is different from the rest? Do a complete hypothesis test. Use $\alpha = 0.05$. Is the post-hoc test necessary for this problem? How many post hoc test would you do? Run all the post hoc tests using Bonferroni approach. What is your conclusion on post-hoc tests?

```r
winston <- c(65,59,57,58,56,59,62,57,57,56)
marija <- c(63,65,59,64,62,68,63,66,65,72)
raiyan <- c(60,56,59,53,56,52,56,58,62,50)
valerie <- c(57,56,59,56,58,60,59,56,62,61)
combined_groups <- data.frame(cbind(winston,marija,raiyan,valerie))
summary(combined_groups)
```

```
##     winston         marija          raiyan         valerie
##  Min.   :56.0   Min.   :59.00   Min.   :50.00   Min.   :56.00
##  1st Qu.:57.0   1st Qu.:63.00   1st Qu.:53.75   1st Qu.:56.25
##  Median :57.5   Median :64.50   Median :56.00   Median :58.50
##  Mean   :58.6   Mean   :64.70   Mean   :56.20   Mean   :58.40
##  3rd Qu.:59.0   3rd Qu.:65.75   3rd Qu.:58.75   3rd Qu.:59.75
##  Max.   :65.0   Max.   :72.00   Max.   :62.00   Max.   :62.00
```

```r
sapply(combined_groups,sd,na.rm=TRUE)
```

```
##  winston   marija   raiyan  valerie
## 2.875181 3.529243 3.735714 2.170509
```

```r
stacked_groups <- stack(combined_groups)
anova_results <- aov(values ~ ind, data= stacked_groups)
summary(anova_results)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## ind          3  399.5  133.16   13.52 4.59e-06 ***
## Residuals   36  354.5    9.85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
pairwise.t.test(stacked_groups$values,stacked_groups$ind,p.adjust.method = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  stacked_groups$values and stacked_groups$ind
##
##         winston marija  raiyan
## marija  0.00065 -       -
## raiyan  0.57505 3.5e-06 -
## valerie 1.00000 0.00042 0.75426
##
## P value adjustment method: bonferroni
```

```r
choose(4,2)
```

```
## [1] 6
```

The p-value for the Anova is significant to determine if we reject the null hypothesis or not, and to determine if there is some difference between them.\ There is a total of 6 post-hoc tests $\binom{4}{2}$.\ The 3 by 3 table given in the computation is the difference of p-value between each variable/name. Depending on the significant difference between variables we can determine to either reject the null hypothesis or not.\ Conclusion of post-hoc tests, is the difference of p-values between: Winston, Marija; Raiyan, Marija; Valerie, Marija. Since the p-values are smaller than α.

4. (20 points) To compare two programs for training industrial workers to perform a skilled job, 20 workers are included in an experiment. Of these, 10 are selected at random and trained by method 1; the remaining 10 are trained by method 2. After completion of training, all the workers are subjected to a time-and-motion test that records the speed of performance of a skilled job. The following time, as measured in minutes, is obtained.

Method										
	Method 1	15	20	11	23	16	21	18	16	27	24
	Method 2	23	31	13	19	23	17	28	26	25	28

Test the hypothesis that the mean job time is equal before and after training with method 1 and 2 versus the alternative that it is significantly less after training with method 1 than after training with method 2. Use a significance level of $\alpha = 0.05$. Assume equal variances.

```r
NullHypothesis <- function(tcrit,tstat)
{if ((abs(tcrit)) > (abs(tstat)))
    {    return ("Not enough information to reject null hypothesis")
    }else
    {return ("Reject null hypothesis")}}

t.test2 <- function(x1,x2,m0,alpha,alternative =
                        c("two.sided", "less", "greater"),equal.variance=FALSE)
{   x=mean(x1)
    y=mean(x2)
    s1=sd(x1)
    s2=sd(x2)
    n1=length(x1)
    n2=length(x2)
    if( equal.variance==FALSE )
    {se <- sqrt( (s1^2/n1) + (s2^2/n2) )
        # welch-satterthwaite df
        df <- ( (s1^2/n1 + s2^2/n2)^2 )/( (s1^2/n1)^2/(n1-1) +
                                              (s2^2/n2)^2/(n2-1) )
    } else
    {# pooled standard deviation, scaled by the sample sizes
        se <- sqrt( (1/n1 + 1/n2) * ((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2) )
        df <- n1+n2-2}
    tstat <- (x-y-m0)/se
    LB <- x-y-abs(qt(alpha/2,df))*se
    UB <- x-y+abs(qt(alpha/2,df))*se
    tcrit <- qt(0.05, df)
    if(alternative=="two.sided")
    {dat <- c(x-y, se, tstat, df, LB,UB,tcrit,
              NullHypothesis(tcrit,tstat))}
    else
    {dat <- c(x-y, se, tstat, df ,LB,UB,tcrit,alpha,
              NullHypothesis(tcrit,tstat))}
    names(dat) <- c("Difference of means", "Std Error", "tstat", "df",
                    "LB","UB","tcrit", "alpha", "Nullhypothesis")
    return(dat)}
t.test2(c(15,20,11,23,16,21,18,16,27,24),
                  c(23,31,13,19,23,17,28,26,25,28),0,0.05, alternative = "less",
                  equal.variance=TRUE)
```

```
##      Difference of means               Std Error                   tstat
##                   "-4.2"       "2.32617950964905"      "-1.80553563582616"
##                       df                       LB                      UB
##                     "18"      "-9.08712180137879"      "0.687121801378789"
##                    tcrit                    alpha           Nullhypothesis
##      "-1.73406360661754"                  "0.05" "Reject null hypothesis"
```

Therefore, $|t_{stat}| > |t_{crit}|$ then we reject the null hypothesis. In other words, there is a strong indication that it is significantly less after training with method 1 than after training with method 2.