# Assignment 3

Grisham Nathan 1001314927

September 27, 2017

# 6.

# A)

library(tidyverse)

## Loading tidyverse: ggplot2  
## Loading tidyverse: tibble  
## Loading tidyverse: tidyr  
## Loading tidyverse: readr  
## Loading tidyverse: purrr  
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats  
## lag(): dplyr, stats

**Code**

myurl = "http://www.utsc.utoronto.ca/~butler/c32/hg.txt"   
input = read\_delim(myurl, " ")

## Parsed with column specification:  
## cols(  
## name = col\_character(),  
## density = col\_double()  
## )

Input

**Output**

## # A tibble: 13 x 2  
## name density  
## <chr> <dbl>  
## 1 jeidji 17.00  
## 2 kuku 50.00  
## 3 mamu 45.00  
## 4 ngatjan 59.80  
## 5 undanbi 21.74  
## 6 jinibarra 16.00  
## 7 ualaria 9.00  
## 8 barkindji 15.43  
## 9 wongaibon 5.12  
## 10 jaralde 40.00  
## 11 tjapwurong 35.00  
## 12 tasmanians 13.35  
## 13 badjalang 13.40

**Explanation**

I have the correct variables. Just as used as column names in the hg.txt, the variables are name and density. 13 hunter-gatherer societies in Australia were studied, as there are 13 rows.

# B)

**Explanation**

Let mu be the mean Let h0 be the null hypothesis Let hA be the alternative hypothesis

h0: mu = 7.38, the mean population density of the societies matches the rest of the world.

hA: mu 7.38, the mean population density is not the same as the rest of the world.

# C)

**Code**

t.test(input$density, mu=7.38)

**Output**

##   
## One Sample t-test  
##   
## data: input$density  
## t = 3.8627, df = 12, p-value = 0.002257  
## alternative hypothesis: true mean is not equal to 7.38  
## 95 percent confidence interval:  
## 15.59244 36.84449  
## sample estimates:  
## mean of x   
## 26.21846

**Explanation**

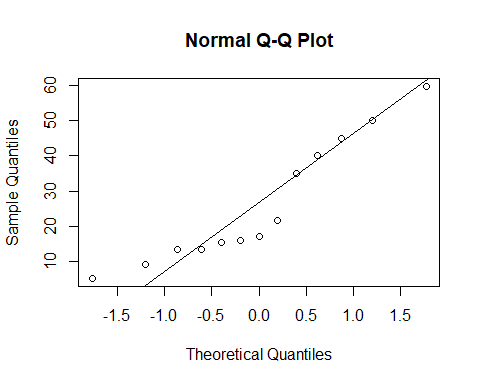
Using one sample satterthwaite t-test, the P-value is 0.002257. Therefore, P-value is lower than 0.05 and the null hypothesis will be rejected.

# D)

**Code**

qqnorm(input$density)  
qqline(input$density)

**Output**



**Explanation**

One of the key assumptions of the t-test is that the data is normally distributed. In addition, there are points that deviate from the theoretical normal distribution line on the Q-Q plot, which show that the data is not being normally distributed. As a result, the t-test that was done is not valid.

# 7.

# A)

**Code**

filename myurl url "http://www.utsc.utoronto.ca/~butler/c32/to-school.csv";

proc import

datafile=myurl

dbms=csv

out=mydata

replace;

getnames=yes;

proc print;

**Output**

| **Obs** | **traveltime** | **location** |
| --- | --- | --- |
| **1** | 30 | Ontario |
| **2** | 10 | Ontario |
| **3** | 8 | Ontario |
| **4** | 30 | Ontario |
| **5** | 5 | Ontario |
| **6** | 8 | Ontario |
| **7** | 7 | Ontario |
| **8** | 15 | Ontario |
| **9** | 10 | Ontario |
| **10** | 35 | Ontario |
| **11** | 15 | Ontario |
| **12** | 10 | Ontario |
| **13** | 25 | Ontario |
| **14** | 22 | Ontario |
| **15** | 20 | Ontario |
| **16** | 25 | Ontario |
| **17** | 30 | Ontario |
| **18** | 10 | Ontario |
| **19** | 25 | Ontario |
| **20** | 8 | Ontario |
| **21** | 15 | Ontario |
| **22** | 18 | Ontario |
| **23** | 25 | Ontario |
| **24** | 15 | Ontario |
| **25** | 10 | Ontario |
| **26** | 25 | Ontario |
| **27** | 5 | Ontario |
| **28** | 2 | Ontario |
| **29** | 5 | Ontario |
| **30** | 25 | Ontario |
| **31** | 20 | Ontario |
| **32** | 15 | Ontario |
| **33** | 47 | Ontario |
| **34** | 20 | Ontario |
| **35** | 20 | Ontario |
| **36** | 13 | Ontario |
| **37** | 20 | Ontario |
| **38** | 5 | Ontario |
| **39** | 15 | Ontario |
| **40** | 12 | Ontario |
| **41** | 45 | UK |
| **42** | 5 | UK |
| **43** | 4 | UK |
| **44** | 15 | UK |
| **45** | 50 | UK |
| **46** | 20 | UK |
| **47** | 20 | UK |
| **48** | 20 | UK |
| **49** | 20 | UK |
| **50** | 20 | UK |
| **51** | 25 | UK |
| **52** | 35 | UK |
| **53** | 15 | UK |
| **54** | 30 | UK |
| **55** | 20 | UK |
| **56** | 10 | UK |
| **57** | 45 | UK |
| **58** | 10 | UK |
| **59** | 3 | UK |
| **60** | 60 | UK |
| **61** | 25 | UK |
| **62** | 20 | UK |
| **63** | 5 | UK |
| **64** | 15 | UK |
| **65** | 5 | UK |
| **66** | 15 | UK |
| **67** | 17 | UK |
| **68** | 30 | UK |
| **69** | 40 | UK |
| **70** | 20 | UK |
| **71** | 10 | UK |
| **72** | 30 | UK |
| **73** | 10 | UK |
| **74** | 15 | UK |
| **75** | 20 | UK |
| **76** | 10 | UK |
| **77** | 15 | UK |
| **78** | 17 | UK |
| **79** | 10 | UK |
| **80** | 25 | UK |

**Code**

proc means mean;

var traveltime;

class location;

**Output**

| **The MEANS Procedure** | | |
| --- | --- | --- |
| **Analysis Variable : traveltime** | | |
| **location** | **N Obs** | **Mean** |
| Ontario | 40 | 17.0000000 |
| UK | 40 | 20.6500000 |

**Explanation**

The mean travel time of Ontario is 17 minutes.

The mean travel time of UK is 20.65 minutes.

There are 40 travel times for each location.

# B)

**Code**

proc sgplot;

vbox traveltime / category=location;

**Output**



**Explanation**

One categorical variable and one quantitative variable is used. Therefore, side by side boxplot will be used. Travel times for UK are more skewed to the right than travel times for Ontario, since UK has outliers. Though, Ontario and UK are both skewed to the right as their quartiles tend to be much closer to their left tails.

UK travel times is more spread out than Ontario travel times, since the standard deviation for Ontario is lower than for the standard deviation of the UK. This is pretty evident from the outliers on the UK boxplot, which will definitely pull the standard deviation of UK up.

# C)

**Code**

proc ttest;

var traveltime;

class location;

**Output**

|  |
| --- |
| ***The TTEST Procedure*** |

|  |
| --- |
| ***Variable: traveltime*** |

| **location** | **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- |
| **Ontario** | 40 | 17.0000 | 9.6609 | 1.5275 | 2.0000 | 47.0000 |
| **UK** | 40 | 20.6500 | 13.1277 | 2.0757 | 3.0000 | 60.0000 |
| **Diff (1-2)** |  | -3.6500 | 11.5254 | 2.5772 |  |  |

| **location** | **Method** | **Mean** | **95% CL Mean** | | **Std Dev** | **95% CL Std Dev** | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ontario** |  | 17.0000 | 13.9103 | 20.0897 | 9.6609 | 7.9138 | 12.4050 |
| **UK** |  | 20.6500 | 16.4516 | 24.8484 | 13.1277 | 10.7537 | 16.8564 |
| **Diff (1-2)** | **Pooled** | -3.6500 | -8.7807 | 1.4807 | 11.5254 | 9.9662 | 13.6676 |
| **Diff (1-2)** | **Satterthwaite** | -3.6500 | -8.7879 | 1.4879 |  |  |  |

| **Method** | **Variances** | **DF** | **t Value** | **Pr > |t|** |
| --- | --- | --- | --- | --- |
| **Pooled** | Equal | 78 | -1.42 | 0.1607 |
| **Satterthwaite** | Unequal | 71.663 | -1.42 | 0.1610 |

| **Equality of Variances** | | | | |
| --- | --- | --- | --- | --- |
| **Method** | **Num DF** | **Den DF** | **F Value** | **Pr > F** |
| **Folded F** | 39 | 39 | 1.85 | 0.0590 |



**Explanation**

Yes, there is evidence of difference between mean travel time between the two locations from the value in the table of means, -3.65. This however is a difference of sample mean travel time, so it will not reflect the difference in mean travel time for the population. However, from the Satterthwaite t-test I conducted, I saw that the p-value is higher than 0.05. As a result, we cannot reject the null hypothesis of mean travel times between the two locations are equal. I used Satterthwaite t-test since the two locations have different standard deviations and therefore unequal variances.

# D)

Yes, I do have concerns about the t-test I just did. The points on the Q-Q plot deviate from the theoretical normal distribution line, as a result we can see that it’s not normal in either location. This means that t-test that was just done, is not valid.