Assignment #5

Grisham Nathan

October 18, 2017

## 8.

## a)

filename myurl url

"http://www.utsc.utoronto.ca/~butler/c32/sunspots.txt";

proc import

datafile=myurl

dbms=dlm

out=sunspots

replace;

delimiter=' ';

getnames=yes;

proc print data=sunspots;

proc means mean median;

var spots;

| **Obs** | **spots** |
| --- | --- |
| **1** | 12.5 |
| **2** | 14.1 |
| **3** | 37.6 |
| **4** | 48.3 |
| **5** | 67.3 |
| **6** | 70 |
| **7** | 43.8 |
| **8** | 56.5 |
| **9** | 59.7 |
| **10** | 24 |
| **11** | 12 |
| **12** | 27.4 |
| **13** | 53.5 |
| **14** | 73.9 |
| **15** | 104 |
| **16** | 54.6 |
| **17** | 4.4 |
| **18** | 177.3 |
| **19** | 70.1 |
| **20** | 54 |
| **21** | 28 |
| **22** | 13 |
| **23** | 6.5 |
| **24** | 134.7 |
| **25** | 114 |
| **26** | 72.7 |
| **27** | 81.2 |
| **28** | 24.1 |
| **29** | 20.4 |
| **30** | 13.3 |
| **31** | 9.4 |
| **32** | 25.7 |
| **33** | 47.8 |
| **34** | 50 |
| **35** | 45.3 |
| **36** | 61 |
| **37** | 39 |
| **38** | 12 |
| **39** | 7.2 |
| **40** | 11.3 |

***The MEANS Procedure***

| **Analysis Variable : spots** | |
| --- | --- |
| **Mean** | **Median** |
| 47.0400000 | 44.5500000 |

The mean is 47.04 and the median is 45.55.

# b)

proc sgplot;

histogram spots;



I think the median would be the better measure of center because the data is skewed to the right. As a result, the mean will be bigger than the median which doesn’t not reflect the fact that most of the data is on the left side. The median will always typically be where the data has most frequency, which in this case is more on the left side of the histogram than the mean.

# c)

proc univariate location=41;

var spots;

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 40 | **Sum Weights** | 40 |
| **Mean** | 47.04 | **Sum Observations** | 1881.6 |
| **Std Deviation** | 37.5228464 | **Variance** | 1407.964 |
| **Skewness** | 1.46537727 | **Kurtosis** | 2.83003603 |
| **Uncorrected SS** | 143421.06 | **Corrected SS** | 54910.596 |
| **Coeff Variation** | 79.7679557 | **Std Error Mean** | 5.93288294 |

| **Basic Statistical Measures** | | | |
| --- | --- | --- | --- |
| **Location** | | **Variability** | |
| **Mean** | 47.04000 | **Std Deviation** | 37.52285 |
| **Median** | 44.55000 | **Variance** | 1408 |
| **Mode** | 12.00000 | **Range** | 172.90000 |
|  |  | **Interquartile Range** | 50.45000 |

| **Tests for Location: Mu0=41** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Student's t** | **t** | 1.018055 | **Pr > |t|** | 0.3149 |
| **Sign** | **M** | 1 | **Pr >= |M|** | 0.8746 |
| **Signed Rank** | **S** | 19 | **Pr >= |S|** | 0.8021 |

| **Quantiles (Definition 5)** | |
| --- | --- |
| **Level** | **Quantile** |
| **100% Max** | 177.30 |
| **99%** | 177.30 |
| **95%** | 124.35 |
| **90%** | 92.60 |
| **75% Q3** | 64.15 |
| **50% Median** | 44.55 |
| **25% Q1** | 13.70 |
| **10%** | 10.35 |
| **5%** | 6.85 |
| **1%** | 4.40 |
| **0% Min** | 4.40 |

| **Extreme Observations** | | | |
| --- | --- | --- | --- |
| **Lowest** | | **Highest** | |
| **Value** | **Obs** | **Value** | **Obs** |
| 4.4 | 17 | 81.2 | 27 |
| 6.5 | 23 | 104.0 | 15 |
| 7.2 | 39 | 114.0 | 25 |
| 9.4 | 31 | 134.7 | 24 |
| 11.3 | 40 | 177.3 | 18 |

# d)

We fail to reject the null hypothesis that median is equal to 41, since sign p-value is 0.8746 which is above significance level of 0.05.

# 

# e)

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

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

## Parsed with column specification:  
## cols(  
## spots = col\_double()  
## )

sunspots %>% slice(1:10)

## # A tibble: 10 x 1  
## spots  
## <dbl>  
## 1 12.5  
## 2 14.1  
## 3 37.6  
## 4 48.3  
## 5 67.3  
## 6 70.0  
## 7 43.8  
## 8 56.5  
## 9 59.7  
## 10 24.0

# f)

sunspots %>% summarize(mean=mean(spots), median=median(spots))

## # A tibble: 1 x 2  
## mean median  
## <dbl> <dbl>  
## 1 47.04 44.55

The mean and median are the same as the ones I calculated in SAS.

# g)

library(devtools)

## Warning: package 'devtools' was built under R version 3.4.2

install\_github("nxskok/smmr")

## Skipping install of 'smmr' from a github remote, the SHA1 (8030a7d1) has not changed since last install.  
## Use `force = TRUE` to force installation

library(smmr)  
sign\_test(sunspots,spots,41)

## $above\_below  
## below above   
## 19 21   
##   
## $p\_values  
## alternative p\_value  
## 1 lower 0.6820860  
## 2 upper 0.4373147  
## 3 two-sided 0.8746293

#or  
#pval\_sign(41,sunspots,spots)

I got the same sign test p-value of 0.8746293 just like in SAS.

# h)

ci\_median(sunspots,spots,conf.level=0.90)

## [1] 25.70298 53.99511

(25.70298, 53.99511) is the 90% confidence interval for the median.

## 9.

# a)

filename myurl url

"http://www.utsc.utoronto.ca/~butler/c32/heights.txt";

proc import

datafile=myurl

dbms=dlm

out=heights

replace;

delimiter=' ';

getnames=yes;

proc print data=heights;

| **Obs** | **reported** | **measured** |
| --- | --- | --- |
| **1** | 67 | 66.3 |
| **2** | 73 | 72.8 |
| **3** | 67.5 | 68 |
| **4** | 70 | 69.4 |
| **5** | 68 | 67.3 |
| **6** | 70 | 70.2 |
| **7** | 71 | 70.9 |
| **8** | 68 | 67.6 |
| **9** | 67.5 | 67.1 |
| **10** | 66 | 65.3 |
| **11** | 70 | 68.8 |
| **12** | 71.5 | 70.9 |

I do have the right rows and columns, as I have noticed that the data is delimited by spaces, which I took account of, when reading the data into SAS. There are 12 rows and 2 columns of data.

# b)

Each of the 12 male students have a height that they each report and the actual height that the experimenters measure. As a result, the heights reported and the heights measured do not each come from two independent samples. The data is matched-pair data, not two independent samples because the data compares two values of height for each male student.

# c)

proc ttest;

paired reported\*measured;

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

|  |
| --- |
| ***Difference: reported - measured*** |

| **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| 12 | 0.4083 | 0.4562 | 0.1317 | -0.5000 | 1.2000 |

| **Mean** | **95% CL Mean** | | **Std Dev** | **95% CL Std Dev** | |
| --- | --- | --- | --- | --- | --- |
| 0.4083 | 0.1185 | 0.6982 | 0.4562 | 0.3232 | 0.7745 |

| **DF** | **t Value** | **Pr > |t|** |
| --- | --- | --- |
| 11 | 3.10 | 0.0101 |









The matched-pairs t-test should be two-sided because our null hypothesis will be that reported heights mean and actual heights mean are equal and our alternative hypothesis will be that reported heights mean and actual heights mean are not equal.

# d)

We reject the null hypothesis of the reported heights mean being equal to actual heights mean, since sign p-value is 0.0101 which is below significance level of 0.05.

# e)

library(tidyverse)  
url="http://www.utsc.utoronto.ca/~butler/c32/heights.txt"  
heights=read\_delim(url," ")

## Parsed with column specification:  
## cols(  
## reported = col\_double(),  
## measured = col\_double()  
## )

heights

## # A tibble: 12 x 2  
## reported measured  
## <dbl> <dbl>  
## 1 67.0 66.3  
## 2 73.0 72.8  
## 3 67.5 68.0  
## 4 70.0 69.4  
## 5 68.0 67.3  
## 6 70.0 70.2  
## 7 71.0 70.9  
## 8 68.0 67.6  
## 9 67.5 67.1  
## 10 66.0 65.3  
## 11 70.0 68.8  
## 12 71.5 70.9

# f)

with(heights,t.test(reported,measured,paired=T))

##   
## Paired t-test  
##   
## data: reported and measured  
## t = 3.1007, df = 11, p-value = 0.01009  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.1184863 0.6981804  
## sample estimates:  
## mean of the differences   
## 0.4083333

In SAS, I got a p-value of 0.0101, however in R, I got a p-value of 0.01009. I believe the p-value of SAS has been rounded up to nearest 4 decimals. As a result, the p-values are definitely the same.