**CSC 423**

**Project 1**

**Joseph Berkowitz**

**Part A. Univariate Data Analysis**

1. **Read in the data from**[**paper1.txt**](http://facweb.cdm.depaul.edu/sjost/csc423/projects/paper1.txt)**. and print it to verify that everything was input correctly.**

pap1 <- read.table("c:/DataSets/paper1.txt", header = T)

print(pap1)

> pap1 <- read.table("c:/DataSets/paper1.txt", header = T)

> print(pap1)

measurer A B

1 1 0.111 0.083

2 2 0.135 0.082

3 3 0.183 0.087

4 4 0.107 0.079

5 5 0.104 0.080

6 6 0.105 0.085

7 7 0.101 0.076

8 8 0.105 0.079

9 9 0.107 0.089

10 10 0.099 0.079

11 11 0.110 0.092

12 12 0.105 0.074

13 13 0.104 0.083

14 14 0.110 0.080

15 15 0.103 0.078

16 16 0.101 0.083

17 17 0.105 0.077

18 18 0.106 0.079

19 19 0.103 0.079

20 20 0.108 0.080

21 21 0.106 0.076

22 22 0.108 0.076

1. **Read in the data from**[**paper2.txt**](http://facweb.cdm.depaul.edu/sjost/csc423/projects/paper2.txt)**and print it to verify that everything was input correctly.**

pap2 <- read.table("c:/DataSets/paper2.txt", header = T)

print(pap2)

> pap2 <- read.table("c:/DataSets/paper2.txt", header = T)

> print(pap2)

measurer brand thickness

1 1 A 0.111

2 1 B 0.083

3 2 A 0.135

4 2 B 0.082

5 3 A 0.183

6 3 B 0.087

7 4 A 0.107

8 4 B 0.079

9 5 A 0.104

10 5 B 0.080

11 6 A 0.105

12 6 B 0.085

13 7 A 0.101

14 7 B 0.076

15 8 A 0.105

16 8 B 0.079

17 9 A 0.107

18 9 B 0.089

19 10 A 0.099

20 10 B 0.079

21 11 A 0.110

22 11 B 0.092

23 12 A 0.105

24 12 B 0.074

25 13 A 0.104

26 13 B 0.083

27 14 A 0.110

28 14 B 0.080

29 15 A 0.103

30 15 B 0.078

31 16 A 0.101

32 16 B 0.083

33 17 A 0.105

34 17 B 0.077

35 18 A 0.106

36 18 B 0.079

37 19 A 0.103

38 19 B 0.079

39 20 A 0.108

40 20 B 0.080

1. **Obtain these univariate statistics separately by brand for the paper thicknesses: sample mean, sample standard deviation, sample median, sample IQR, these percentiles: 5, 10, 25, 75, 90, 95.  You can use the SAS proc means or proc univariate to compute these statistics. Don't compute them by hand.**

Sample Mean of Brand A = 0.1102727

Sample Mean of Brand B = 0.08072727

Sample Standard Deviation of Brand A = 0.01768018

Sample Standard Deviation of Brand B = 0.004474072

Sample Median of Brand A = 0.1055

Sample Median of Brand B = 0.0795

Sample IQR of Brand A = 0.004

Sample IQR of Brand B = 0.00475

Quantiles of Brand A:

5% 10% 25% 75% 90% 95%

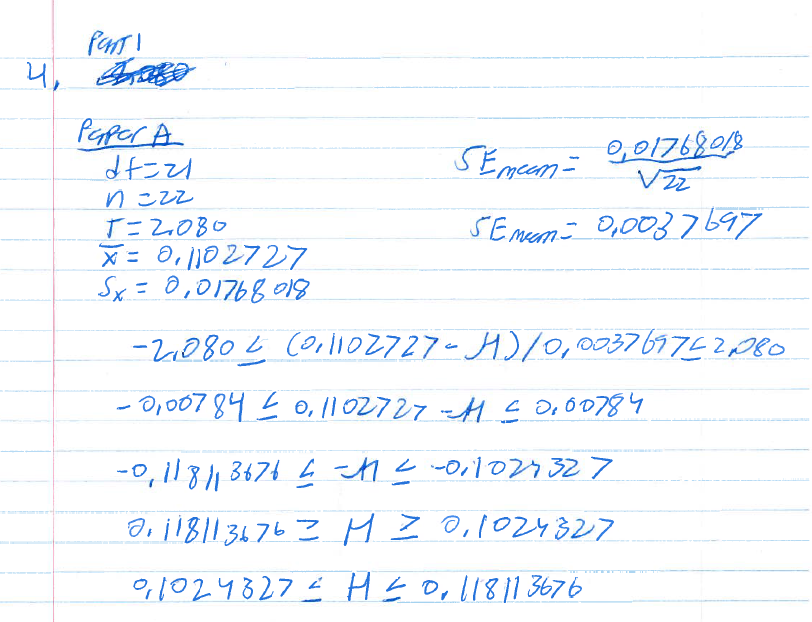
0.1010 0.1012 0.1040 0.1080 0.1109 0.1338

Quantiles of Brand B

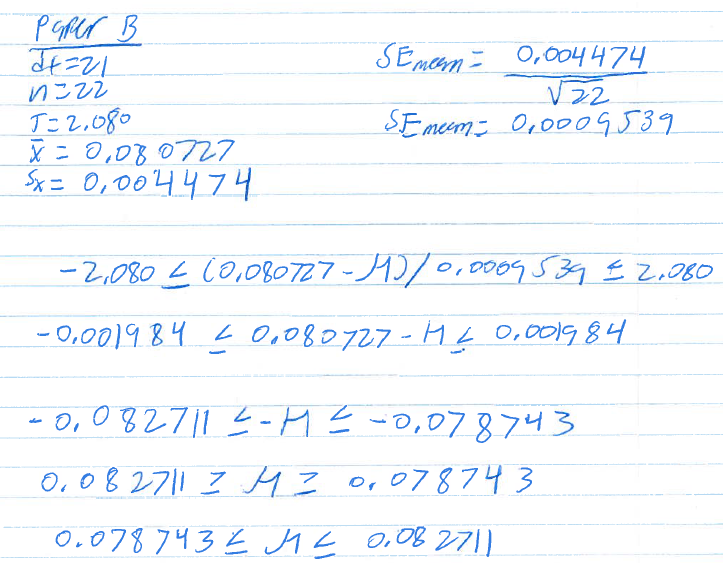
5% 10% 25% 75% 90% 95%

0.07600 0.07600 0.07825 0.08300 0.08680 0.08890

1. **Find 95% confidence intervals for the true thickness for each type of paper separately.  Show your hand calculations and also show the relevent SAS or R output to verify your calculations.**

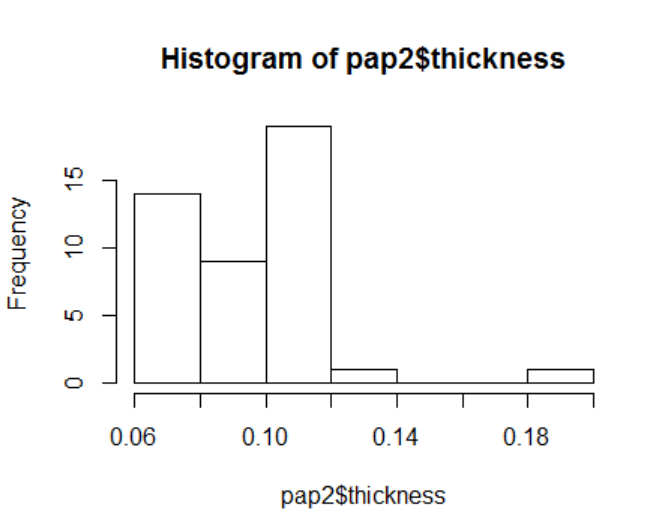


Based on the calculations above, we are 95% confident that the true thickness of paper A is between 0.102431597 and 0.118113803 mm.

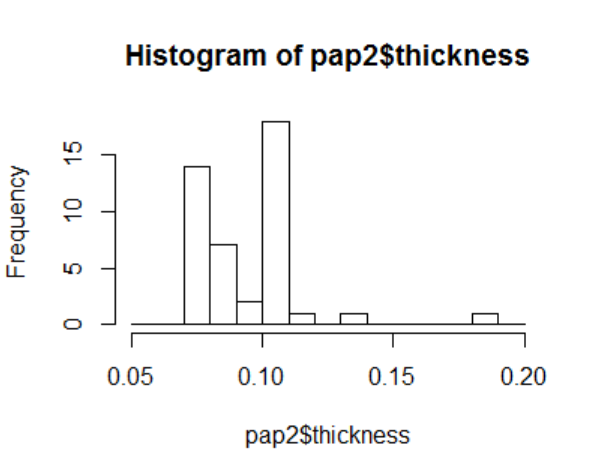


Based on the calculations above, we are 95% confident that the true thickness of paper B is between 0.07828846 and 0.0822569366 mm.

1. **Create three histograms for thicknesses from the combined types of paper:**
   1. **Create a histogram using the default setting for the number of bins  Run your SAS or R code first without the code in steps 6b or 6c to see what bins are obtained with the default setting.**



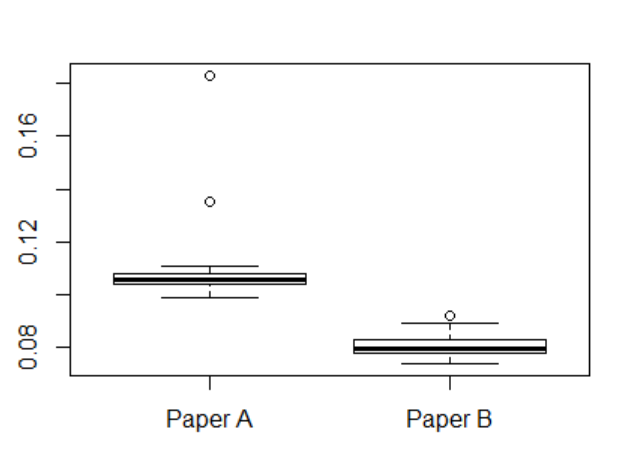
* 1. **Create a histogram with more bins than the default. In SAS, you can do this with an option on the histogram statement of proc univariate**



* 1. **Create a histogram with less bins than the default. (See step 6b.)**



1. **Create side-by-side boxplots of the thicknesses for Brand A and Brand B. Discuss what the boxplots tell you. Are there any outliers? If you are using SAS, use the paper2.txt dataset and sort the dataset by brand before plotting the boxplots.**



The boxplots for the thickness of Paper A and Paper B tell very different things.

First, for paper A, it shows that the distribution of the values is fairly close together. However, there are a couple of outliers at the top end of the distribution, which aren’t included in the boxplot. The values for paper A are slightly skewed toward the top of the distribution, but for the most part the minimum and maximum are fairly close, excluding the outliers.

For paper B, it shows that the distribution of values is a little more spread out, especially between the minimum and maximum. There is one outlier, but it isn’t as far off from the top end of the distribution as the outliers for Paper A.

**Part B. One-sample t-test**

1. **Create a SAS or R dataset containing the number of concurrent users at each location. For example, call your variable nusers. If you are using SAS and copy the preceding data lines verbatim, don't forget the trailing @@ in the input statement. Use this R statement to read data from within the script:**

nusers <- c(scan( ))

17.2 22.1 18.5 17.2 18.6 14.8 21.7 15.8 16.3 22.8

24.1 13.3 16.2 17.5 19.0 23.9 14.8 22.2 21.7 20.7

13.5 15.8 13.1 16.1 21.9 23.9 19.3 12.0 19.9 19.4

15.4 16.7 19.5 16.2 16.9 17.1 20.2 13.4 19.8 17.7

19.7 18.7 17.6 15.9 15.2 17.1 15.0 18.8 21.6 11.9

print(nusers)

> nusers <- c(scan( ))

1: 17.2 22.1 18.5 17.2 18.6 14.8 21.7 15.8 16.3 22.8

11: 24.1 13.3 16.2 17.5 19.0 23.9 14.8 22.2 21.7 20.7

21: 13.5 15.8 13.1 16.1 21.9 23.9 19.3 12.0 19.9 19.4

31: 15.4 16.7 19.5 16.2 16.9 17.1 20.2 13.4 19.8 17.7

41: 19.7 18.7 17.6 15.9 15.2 17.1 15.0 18.8 21.6 11.9

51:

Read 50 items

> nusers

[1] 17.2 22.1 18.5 17.2 18.6 14.8 21.7 15.8 16.3 22.8 24.1 13.3 16.2 17.5 19.0

[16] 23.9 14.8 22.2 21.7 20.7 13.5 15.8 13.1 16.1 21.9 23.9 19.3 12.0 19.9 19.4

[31] 15.4 16.7 19.5 16.2 16.9 17.1 20.2 13.4 19.8 17.7 19.7 18.7 17.6 15.9 15.2

[46] 17.1 15.0 18.8 21.6 11.9

> print(nusers)

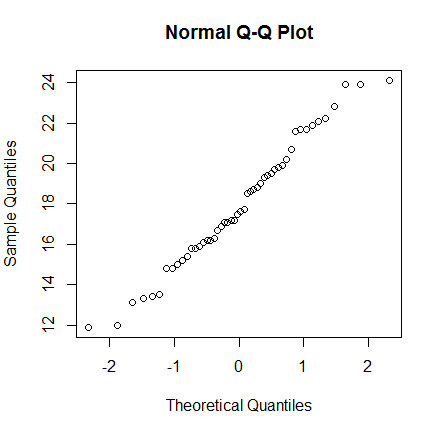
[1] 17.2 22.1 18.5 17.2 18.6 14.8 21.7 15.8 16.3 22.8 24.1 13.3 16.2 17.5 19.0

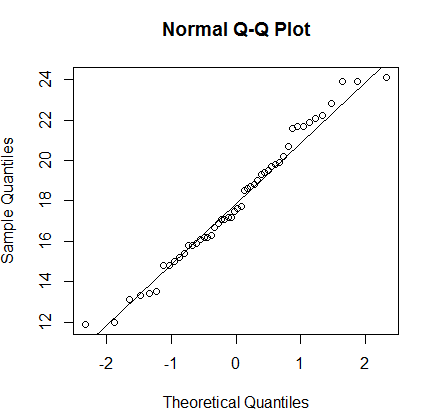
[16] 23.9 14.8 22.2 21.7 20.7 13.5 15.8 13.1 16.1 21.9 23.9 19.3 12.0 19.9 19.4

[31] 15.4 16.7 19.5 16.2 16.9 17.1 20.2 13.4 19.8 17.7 19.7 18.7 17.6 15.9 15.2

[46] 17.1 15.0 18.8 21.6 11.9

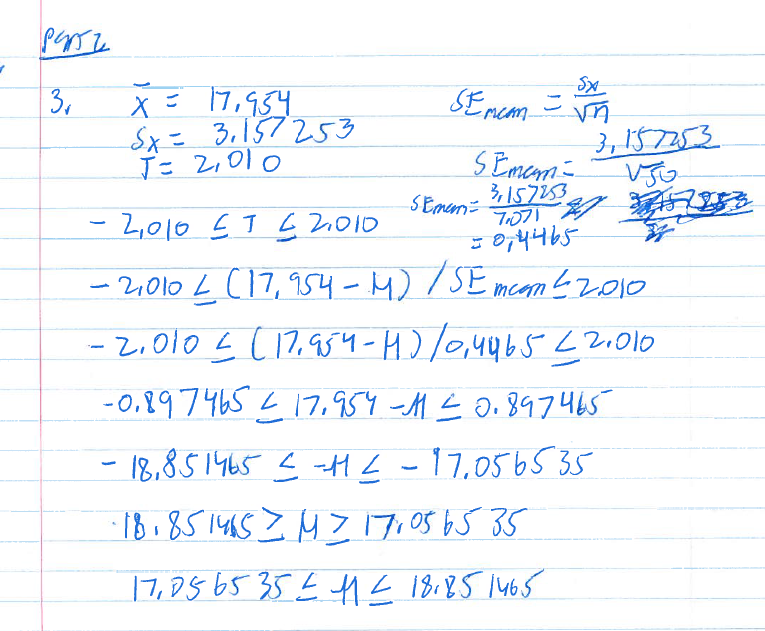
**2. Create create and interpret the normal plot for nusers.**





The plot of nusers looks to be fairly normal. The values are clustered in the middle, and those values near the tails lie on each side of the fit line.

**3. Compute a 95% confidence interval for nusers.  Show your hand calculations with the relevant SAS or R output.  Although standard normal confidence interval [-1.96,1.96] can be used, use the more precise t-distribution interval obtained from the t-table for 50 - 1 = 49 degrees of freedom instead.  You can check your answer with SAS using proc means or proc ttest.  You can check your answer with R using the t.test function.**



t.test(nusers, y=NULL, mu = 17.2)

One Sample t-test

data: nusers

t = 1.6887, df = 49, p-value = 0.09764

alternative hypothesis: true mean is not equal to 17.2

95 percent confidence interval:

17.05672 18.85128

sample estimates:

mean of x

17.954

**4. Show the five steps of the one-sample t-test at the α = 0.05 level to test whether nusers has changed in the past month. Usage data from last month shows an average of 17.2 thousand concurrent users. You don't need to perform hand calculations for steps 1, 2, and 5; just copy the values from the SAS or R output. For Step 3, find the 95% confidence interval using the t-table.**

1. Write down the null and the alternative hypotheses.

H0: true mean of nusers is 17.2

H1: true mean of nusers is not 17.2

1. Compute the test statistic z or t assuming the null hypothesis.

t = 1.6887

1. Determine a 100(1-α)% confidence interval I for the test statistic.

> qt(0.975, 49)

Confidence interval for the test statistic is:

(-2.009575, 2.009575)

1. [1] 2.009575Decide whether to accept or reject the null hypothesis:  
   if t ∈ I, accept H0; if t ∉ I, accept H0;t ∉ I

Accept H0

1. Compute the p-value.

p-value = 0.09764

Part C. Two-sample t-tests (25 pts.)

Use the data in paper1-cleaned.txt and/or paper2-cleaned.txt to answer the following questions. These datasets are the paper1 and paper2 datasets with the outliers removed.

1. **If you are using SAS, create labels for each variable thickness and brand. If you are using R, add print statements in your source code to explain what your output means.**

pap1clean <- read.table("c:/DataSets/paper1-cleaned.txt", header = T)

print(pap1clean)

pap2clean <- read.table("c:/DataSets/paper2-cleaned.txt", header = T)

print(pap2clean)

thickAClean <- c(pap1clean$A)

print(thickAClean)

thickBClean <-c(pap1clean$B)

print(thickBClean)

> thickAClean <- c(pap1clean$A)

> print(thickAClean)

[1] 0.111 0.107 0.104 0.105 0.101 0.105 0.107 0.099 0.105 0.104 0.110 0.103

[13] 0.101 0.105 0.106 0.103 0.108 0.106 0.108

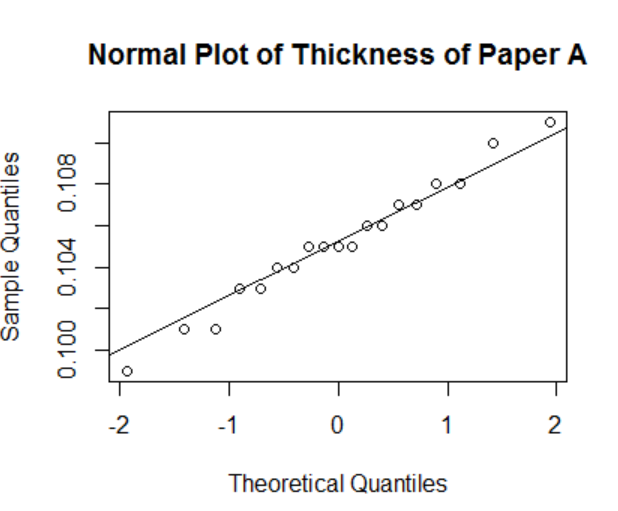
> thickBClean <-c(pap1clean$B)

> print(thickBClean)

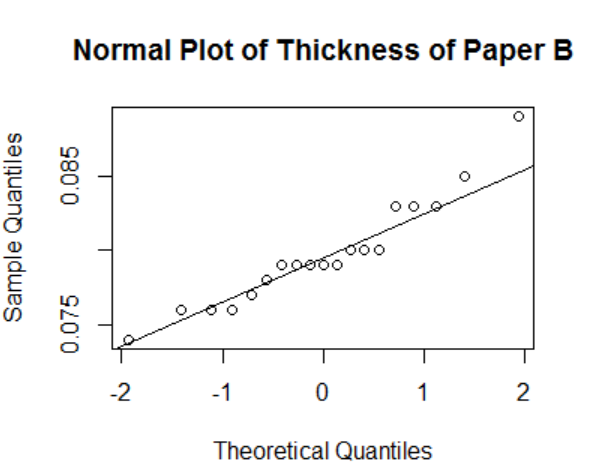
[1] 0.083 0.079 0.080 0.085 0.076 0.079 0.089 0.079 0.074 0.083 0.080 0.078

[13] 0.083 0.077 0.079 0.079 0.080 0.076 0.076

1. Create normal plots of the thicknesses separately for the paper brands A and B. Interpret these normal plots.



This normal plot of Paper A looks as though it may have a fairly normal distribution, with most of the values being on or near the trend line.



The normal plot of Paper B looks like it may have thick tails. The values are on both sides of the trend line.

1. Type out the five steps of a 0.05-level paired-sample t-test to test the null hypothesis that there is no difference between the paper thicknesses in paper1-cleaned.txt. Show relevent SAS or R output in your report. You will need to obtain the confidence interval for the test statistic from the t-table.

> t.test(thickAClean, thickBClean, alternative="two.sided",

+ paired=TRUE, conf.level=0.95)

Paired t-test

data: thickAClean and thickBClean

t = 25.547, df = 18, p-value = 1.357e-15

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.0233305 0.0275116

sample estimates:

mean of the differences

0.02542105

* 1. Write down the null and the alternative hypotheses.

H0: There is no difference in the true paper thicknesses in paper1-cleaned.txt

H1: There is a difference in the true paper thicknesses in paper1-cleaned.txt

* 1. Compute the test statistic z or t assuming the null hypothesis.

t = 25.547

* 1. Determine a 100(1-α)% confidence interval I for the test statistic.

(-2.10092, 2.100922)

* 1. Decide whether to accept or reject the null hypothesis:

We will reject H0, as the test statistic does not fall within the 95% confidence interval for the test statistic.

* 1. Compute the p-value.

P-value = 1.357e-15

1. Type out the five steps a 0.05-level independent two-sample t-test to test the null hypothesis that there is no difference between the paper thicknesses for brands A and B. Show your output and discuss what it means. Use paper2-cleaned.txt sorted by brand for SAS but paper1-cleaned.txt for R. You will need to obtain the confidence interval for the test statistic from the t-table.

> t.test(thickAClean, thickBClean, alternative="two.sided",

+ paired=FALSE, conf.level=0.95)

Welch Two Sample t-test

data: thickAClean and thickBClean

t = 23.5, df = 35.014, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.02322499 0.02761711

sample estimates:

mean of x mean of y

0.10515789 0.07973684

> qt(.975, 35.014)

[1] 2.030079

* 1. Write down the null and the alternative hypotheses.

H0: There is no difference in the true paper thicknesses in paper1-cleaned.txt

H1: There is a difference in the true paper thicknesses in paper1-cleaned.txt

* 1. Compute the test statistic z or t assuming the null hypothesis.

t = 23.5

* 1. Determine a 100(1-α)% confidence interval I for the test statistic.

(-2.030079, 2.030079)

* 1. Decide whether to accept or reject the null hypothesis:

We will reject H0

* 1. Compute the p-value.

p-value < 2.2e-16

1. Is the paired sample or the independent two-sample t-test is more appropriate to decide if the true thickness of a sheet of paper is different for Brand A or Brand B? Explain your answer. How do the p-values compare for the two t-tests? Is this what you would expect?

I believe that the paired sample t-test is the most appropriate test to decide the true thickness. The paired sample test is more appropriate because both brand A and brand B are being measured by the same person. Both of the p-values for the different tests are very small (1.357e-15 for paired sample, p-value < 2.2e-16 for independent sample). Given that each of the test statistics are fairly large in each case, and we reject the null hypothesis in each case, the small p-values are as expected.