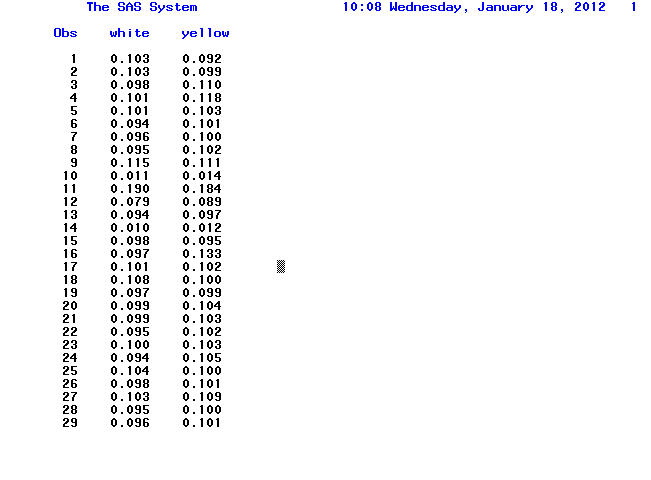
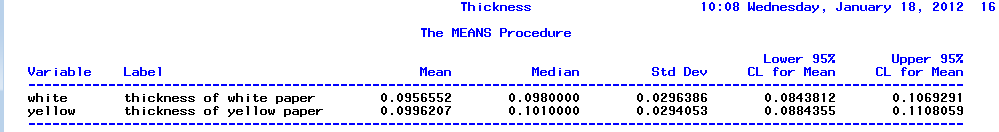
**Part A. Univariate Data Analysis (70 pts.)**

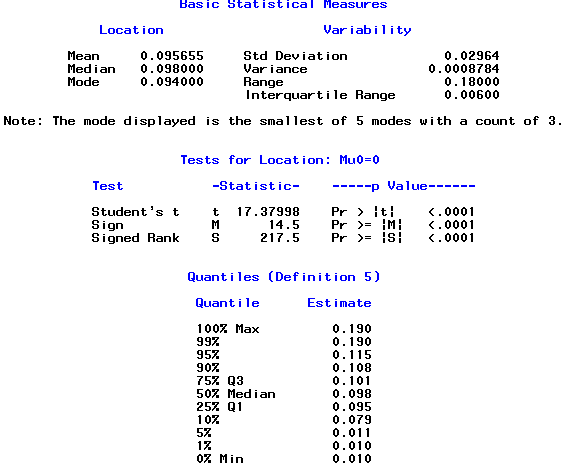
**The measurements in the Paper dataset in** [**paper.txt**](http://facweb.cs.depaul.edu/sjost/csc423/projects/paper.txt) **consist of the paper thicknesses of two brands of paper (White and Yellow) that we measured in class with the micrometer. Use this data to do the following:**

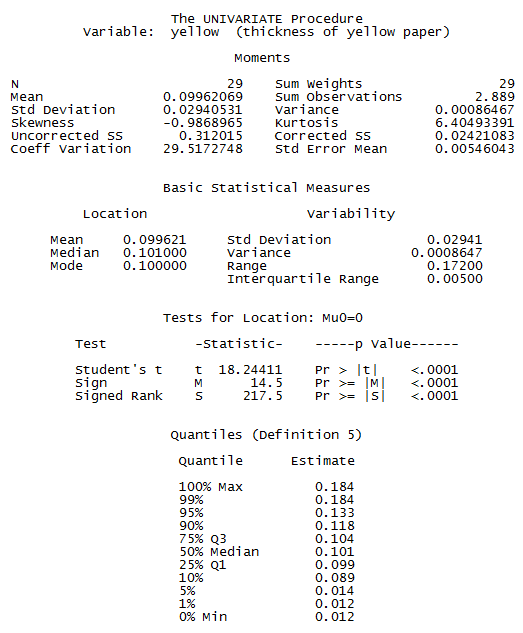
**1. Read in the data from** [**paper.txt**](http://facweb.cs.depaul.edu/sjost/csc423/projects/paper.txt) **and print it to verify that everything was input correctly.**

**2. 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.   The cat statement in R is good for labeling your output.**



**3. 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. Type the answers to this question, but also show the relevant output from SAS or R to justify your answers.**





**White paper:**

sample mean == 0.0956552

sample standard deviation= 0.0296386

sample median=(n+1)/2 ranked value=0.0980000

sample IQR IQR=Q3-Q1=0.006 Q3= ranked value=0.101

Q1= ranked value=0.095

Percentiles:

90% 0.108

75% Q3 0.101

50% Median 0.098

25% Q1 0.095

10% 0.079

5% 0.011

**Yellow paper:**

sample mean == 0.0996207

sample standard deviation= 0.0294053

sample median=(n+1)/2 ranked value= 0.1010000

sample IQR IQR=Q3-Q1=0.005

Q3= ranked value=0.104

Q1= ranked value=0.099

Percentiles:

90% 0.118

75% Q3 0.104

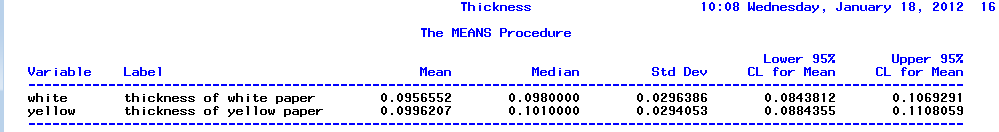
50% Median 0.101

25% Q1 0.099

10% 0.089

5% 0.014

**4. Find 95% confidence intervals for the true thickness for each type of paper separately. Show your work and relevant SAS output.**

****

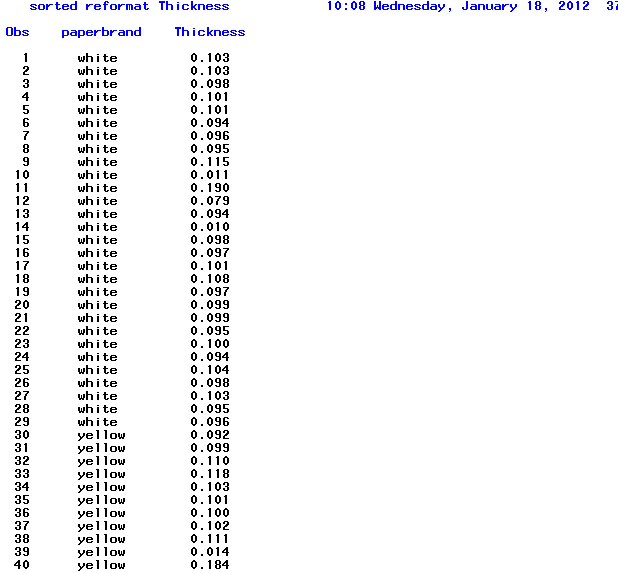
Equation for 95% confidence interval



White paper`s 95% confidence interval is [0.0843812, 0.1069291]

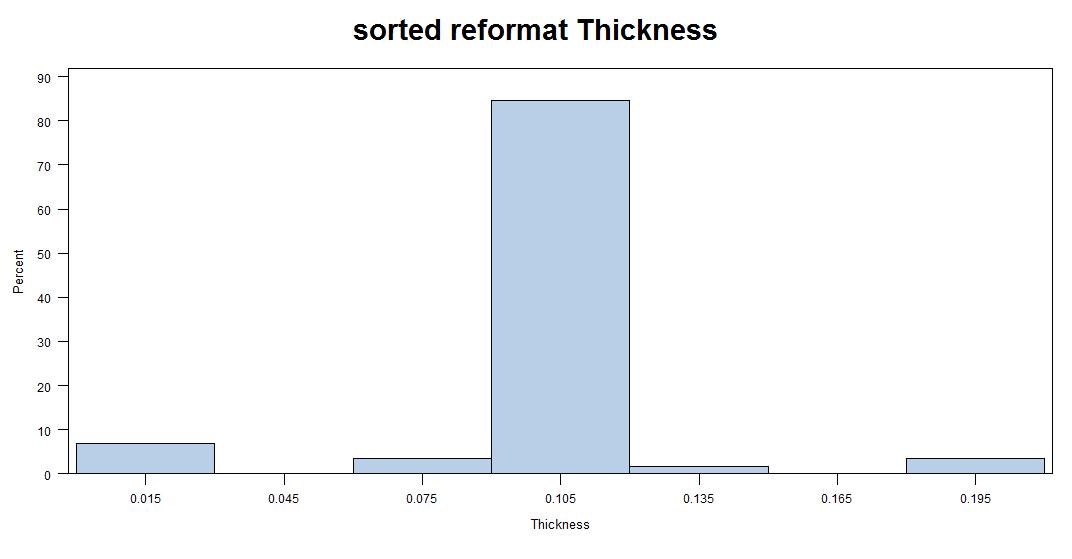
Yellow paper`s 95% confidence interval is [0.0884355, 0.1108059]

**5. Create new dataset where the thicknesses of both types of paper are combined into one variable called "thickness". (See the Reformat Example.) Create a second variable named brand that indicates the type of paper: "White" or "Yellow". See the Reformat Example. Print this new dataset to verify that everything was input correctly. Here are** [**SAS code**](http://facweb.cs.depaul.edu/sjost/csc423/projects/sas-hints.txt) **and** [**R code**](http://facweb.cs.depaul.edu/sjost/csc423/projects/r-hints.txt)**, modified from the Reformat Example to do this.**



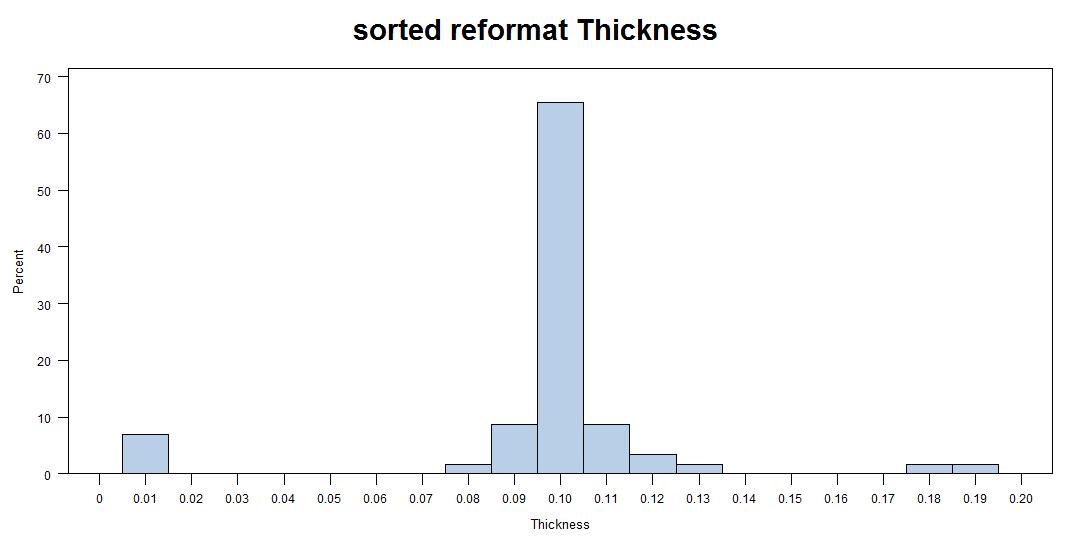
**6. Create three histograms for thicknesses from the combined types of paper:**

**a. 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.**

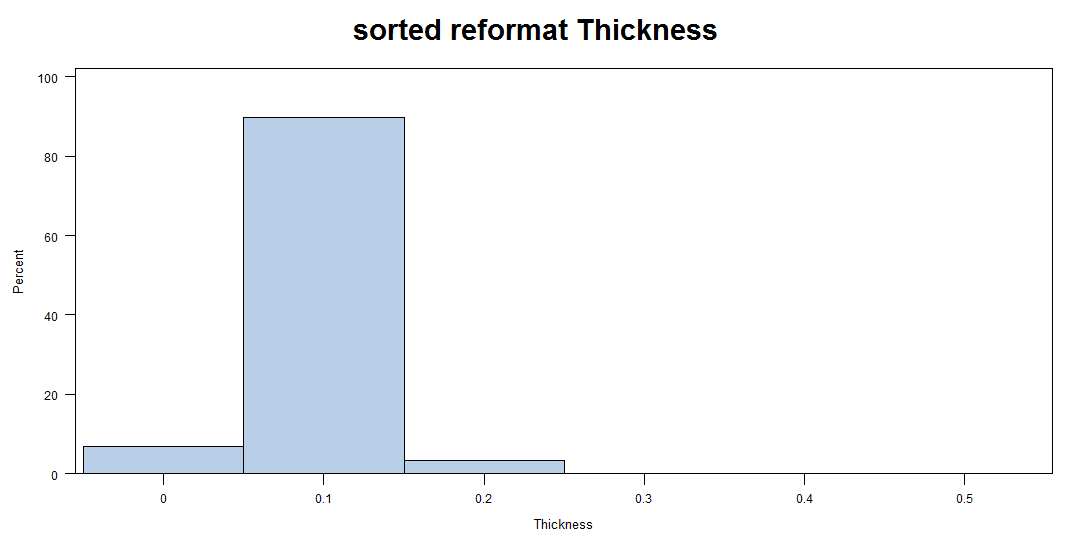
****

b. 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:

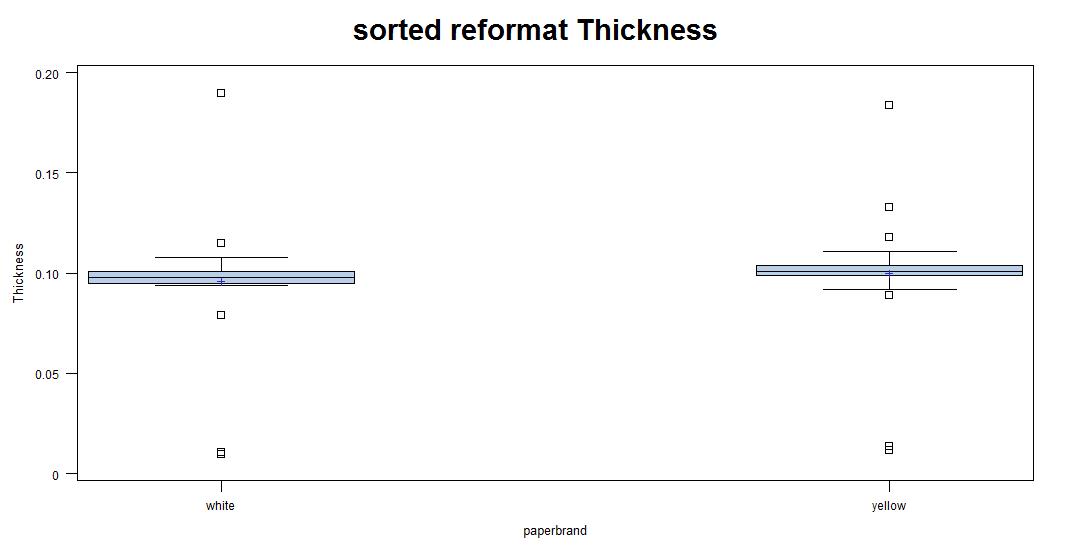
**histogram / midpoints=11 to 25 by 1.0;**

****

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



**7.**

****

* The IQR of yellow is bigger than that of white paper.
* The distributions of two sets are both similar with normal distribution.
* Most samples from yellow are bigger than those from white.
* White has an outlier. Laser has two outliers as denoted by box

### Part B. Short Essay Questions (30 pts.)

### Answer the following questions in complete sentences with paragraphs. Supply an introduction and conclusion where appropriate. Use the textbook or other references if you wish (including the internet), but include a citation in your submission to show where you obtained the information. Answer all questions. Two to three paragraphs each.

### 1. Compare the sample mean and sample median as statistics for estimating the center of a probability density. Discuss the advantages and disadvantages of each.  What are some alternatives to the sample mean and sample median for estimating the center of a probability density?

Answer:

As measures of central tendency, the mean and the median each have advantages and disadvantages.

The mean is the arithmetic average of a set of values, or distribution. It is the most widely used statistic tool to derive the central tendency. It is easy to compute and is less influenced by different sample sets. However, mean is sensitive to extreme scores when population samples are small. It is not a robust tool since it is largely influenced by outliers.

The median is the middle score in a list of scores; it is the point at which half the scores are above and half the scores are below. Medians are less sensitive to extreme scores and are probably a better indicator generally of where the middle of the class is achieving, especially for smaller sample sizes. Also, the median is better suited for skewed distributions to derive at central tendency since it is much more robust and sensible. The disadvantage of median is that it is difficult to handle theoretically. There is no easy mathematical formula to calculate the median[1].

Therefore, when the sample size is large and does not include outliers, the mean score usually provides a better measure of central tendency. The median may be a better indicator of the most typical value if a set of scores has an outlier. Use the mean if you know that the distribution is approximately normal. Use the median if your dataset has extreme outliers, or is skewed to the left or right.(reference: www.google.com)

### 2. Why is the normal distribution so popular in statistics?

Answer: The normal distribution is one which appears in a variety of statistical applications. One reason for this is the central limit theorem. The CLT states that even if independent observations from a population do not have a normal distribution, the sample mean of the observations is approximately normally distributed if n is large. So normal distribution is widely used in many areas of statistics.

A second reason the normal distribution is so important is that it is easy for mathematical statisticians to work with. This means that many kinds of statistical tests can be derived for normal distributions. Almost all statistical tests discussed in this text assume normal distributions. Fortunately, these tests work very well even if the distribution is only approximately normally distributed. Some tests work well even with very wide deviations from normality.  
  
Finally, if the [mean](http://davidmlane.com/hyperstat/A15885.html) and [standard deviation](http://davidmlane.com/hyperstat/A16252.html) of a normal distribution are known, it is easy to convert back and forth from raw scores to percentiles.

**3. Explain the difference among the R functions dnorm, pnorm, qnorm, rnorm. Give examples and explain how they are useful.**

Answer:

Dnorm gives the density, pnorm gives the distribution function, qnorm gives the quantile function, and rnorm generates random deviates.

> dnorm(0)

[1] 0.3989423

> dnorm(0)\*sqrt(2\*pi)

[1] 1

> dnorm(0,mean=4)

[1] 0.0001338302

> dnorm(0,mean=4,sd=10)

[1] 0.03682701

>v <- c(0,1,2)

> dnorm(v)

[1] 0.39894228 0.24197072 0.05399097

> x <-seq(-20,20,by=.1)

> y <- dnorm(x)

> plot(x,y)

> y <- dnorm(x,mean=2.5,sd=0.1)

> plot(x,y)

*pnorm* computes the probability that a normally distributed random number will be less than that number. This function also goes by the rather ominous title of the "Cumulative Distribution Function."

> pnorm(0)

[1] 0.5

> pnorm(1)

[1] 0.8413447

> pnorm(0,mean=2)

[1] 0.02275013

> pnorm(0,mean=2,sd=3)

[1] 0.2524925

> v <- c(0,1,2)

> pnorm(v)

[1] 0.5000000 0.8413447 0.9772499

> x <- seq(-20,20,by=.1)

> y <- pnorm(x)

> plot(x,y)

> y <- pnorm(x,mean=3,sd=4)

> plot(x,y)

*qnorm* is the inverse of *pnorm*. The idea behind *qnorm* is that you give it a probability, and it returns the number whose cumulative distribution matches the probability. For example, if you have a normally distributed random variable with mean zero and standard deviation one, then if you give the function a probability it returns the associated Z-score:

> qnorm(0.5)

[1] 0

> qnorm(0.5,mean=1)

[1] 1

> qnorm(0.5,mean=1,sd=2)

[1] 1

> qnorm(0.5,mean=2,sd=2)

[1] 2

> qnorm(0.5,mean=2,sd=4)

[1] 2

> qnorm(0.25,mean=2,sd=2)

[1] 0.6510205

> qnorm(0.333)

[1] -0.4316442

> qnorm(0.333,sd=3)

[1] -1.294933

> qnorm(0.75,mean=5,sd=2)

[1] 6.34898

> v = c(0.1,0.3,0.75)

> qnorm(v)

[1] -1.2815516 -0.5244005 0.6744898

> x <- seq(0,1,by=.05)

> y <- qnorm(x)

> plot(x,y)

> y <- qnorm(x,mean=3,sd=2)

> plot(x,y)

> y <- qnorm(x,mean=3,sd=0.1)

> plot(x,y)

*rnorm* function can generate random numbers whose distribution is normal. The argument that you give it is the number of random numbers that you want, and it has optional arguments to specify the mean and standard deviation:

> rnorm(4)

[1] 1.2387271 -0.2323259 -1.2003081 -1.6718483

> rnorm(4,mean=3)

[1] 2.633080 3.617486 2.038861 2.601933

> rnorm(4,mean=3,sd=3)

[1] 4.580556 2.974903 4.756097 6.395894

> rnorm(4,mean=3,sd=3)

[1] 3.000852 3.714180 10.032021 3.295667

> y <- rnorm(200)

> hist(y)

> y <- rnorm(200,mean=-2)

> hist(y)

> y <- rnorm(200,mean=-2,sd=4)

> hist(y)

> qqnorm(y)

> qqline(y)

**Reference**

[1] http://www.diffen.com/difference/Mean\_vs\_Median

[2] http://davidmlane.com/hyperstat/A25329.html

[3] http://www.cyclismo.org/tutorial/R/probability.html

[4] http://www.google.com