[To Projects](http://facweb.cs.depaul.edu/sjost/csc423/projects.htm" \t "_top)

**CSC 423/324 -- Project 2**

**See the submission specs and other guidelines at the top of the** [**Project 1 Description**](http://facweb.cs.depaul.edu/sjost/csc423/projects/proj1.htm)**.**

**Part A. One-sample t-test (30 pts.)**

**To investigate the load on its network, a technician records the number of concurrent users at fifty locations (thousands of people):**

**17.2 22.1 18.5 17.2 18.6 14.8 21.7 15.8 16.3 22.8**

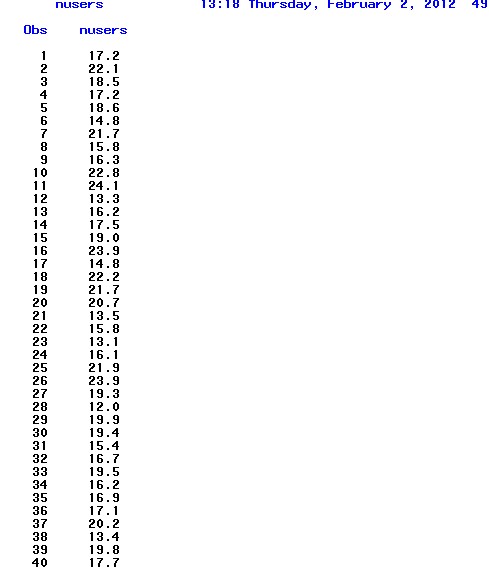
**24.1 13.3 16.2 17.5 19.0 23.9 14.8 22.2 21.7 20.7**

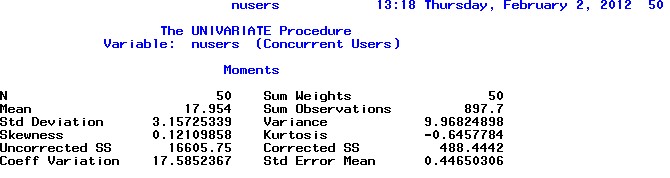
**13.5 15.8 13.1 16.1 21.9 23.9 19.3 12.0 19.9 19.4**

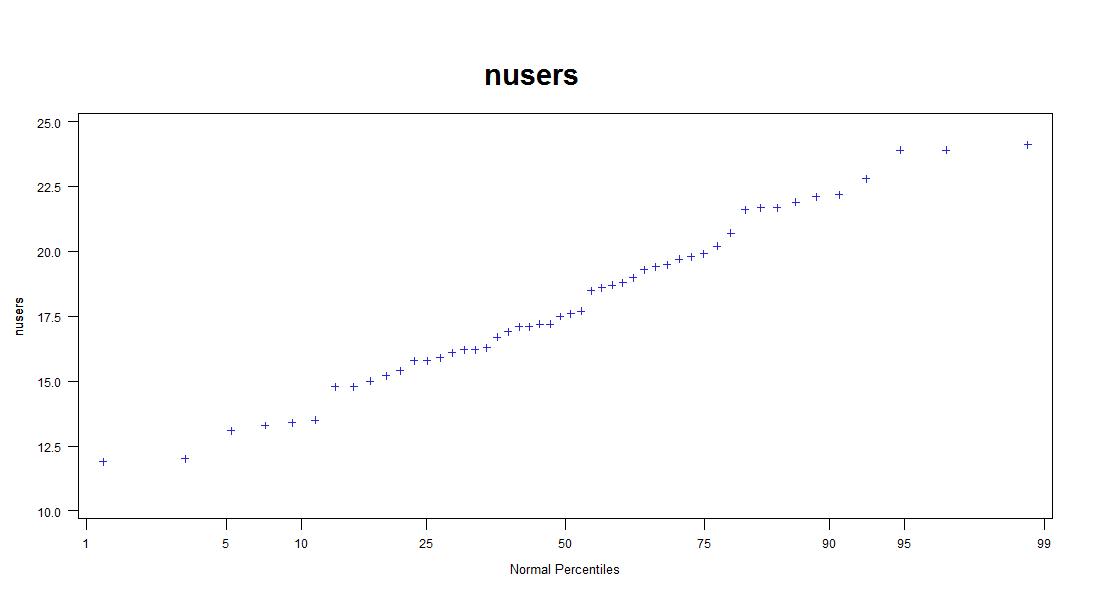
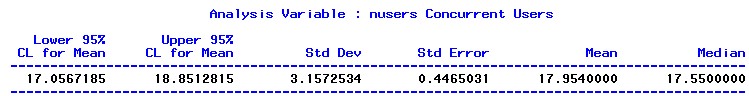
**15.4 16.7 19.5 16.2 16.9 17.1 20.2 13.4 19.8 17.7**

**19.7 18.7 17.6 15.9 15.2 17.1 15.0 18.8 21.6 11.9**

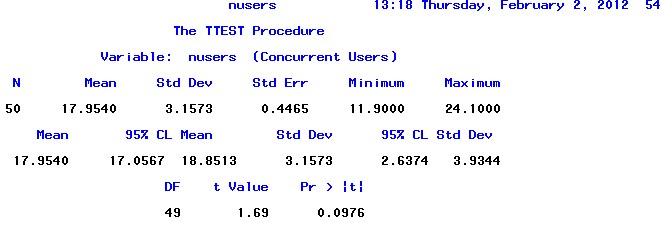
**1. Create a SAS or R dataset containing the number of concurrent users at each location. For example, call your variable nusers.**



**2. If you are using SAS, create a label for nusers.** 

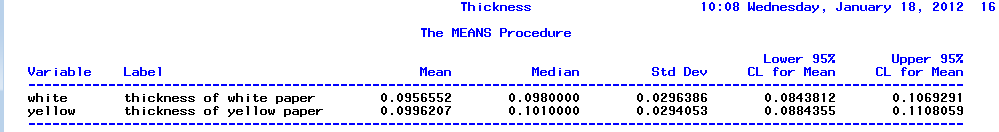
**3. Create a normal plot for nusers.**   
**4. Compute a 95% confidence interval for nusers.  Show your work with the relevant SAS or R output.** 

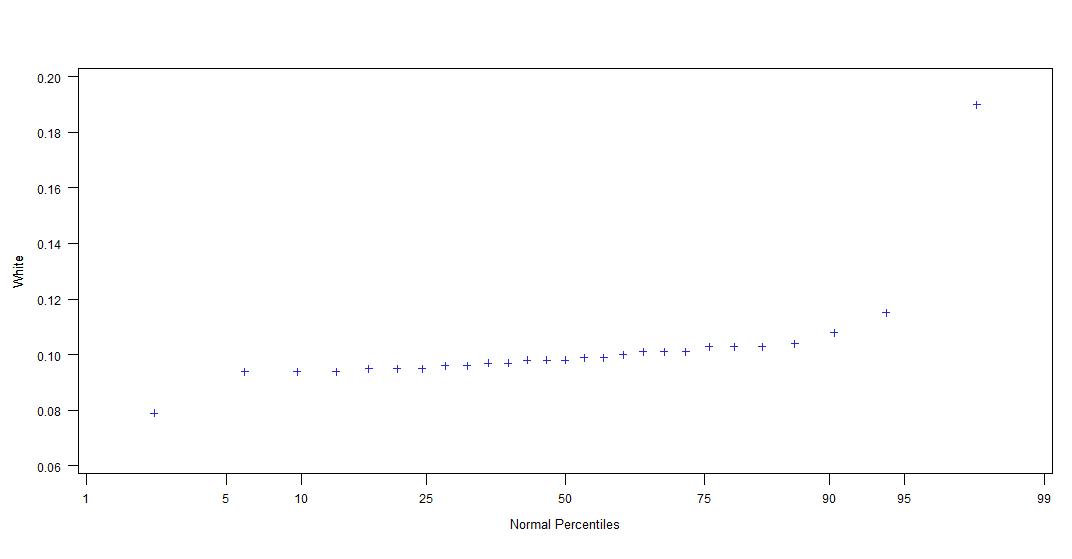
**5. Perform a t-test to test whether the nusers has changed in the past month. Usage data from last month shows an average of 17.2 thousand concurrent users.**

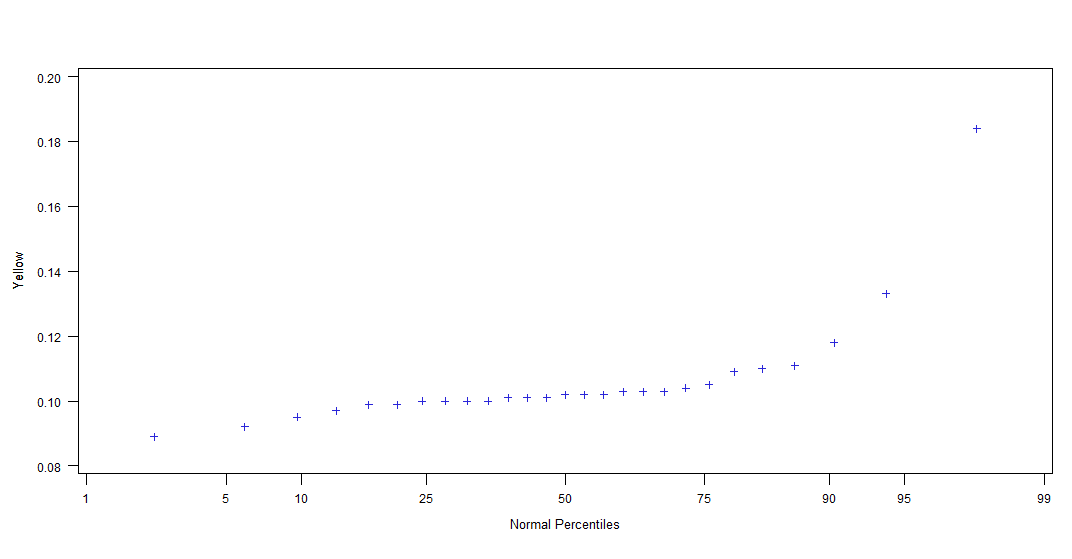


**Part B. Univariate Data Analysis (30 pts.)**

**The use the data in** [**paper.txt**](http://facweb.cs.depaul.edu/sjost/csc423/projects/paper.txt) **that you use for Project 1 for this part. Observations 10 and 14 are clearly incorrectly measured. You can omit them before you do your analyses, or, for 3 points extra credit,  use the subsetting technique in the Subset example to remove the observations less than 0.05 from the dataset.**

**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.**   
 

**2.Create normal plots of the thicknesses separately for the paper brands White and Yellow.  Discuss what the normal plots tell you.**   
 



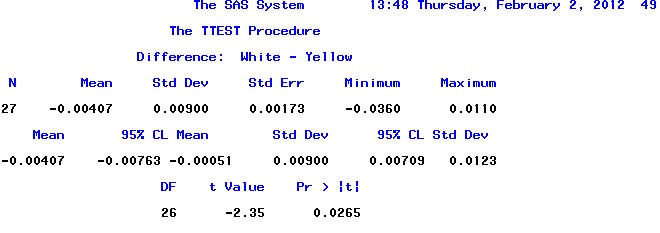
**Answer:**

Generally, data is normally distributed thus can be properly used for t-test, if omit those outliers.

**3. Identify any outliers in the thicknesses for the paper types.  (Outliers are points clearly below the reference line of the normal plot on the left or clearly above the reference line on the right.  Such points are further away from the median than would be expected if the data were normal.)  Justify your answer.**

**Answer:**

According to the box plots, outliers of white papers are 0.19, 0.115,and 0.079. Outliers of yellow paper are 0.184, 0.133 and 0.089.

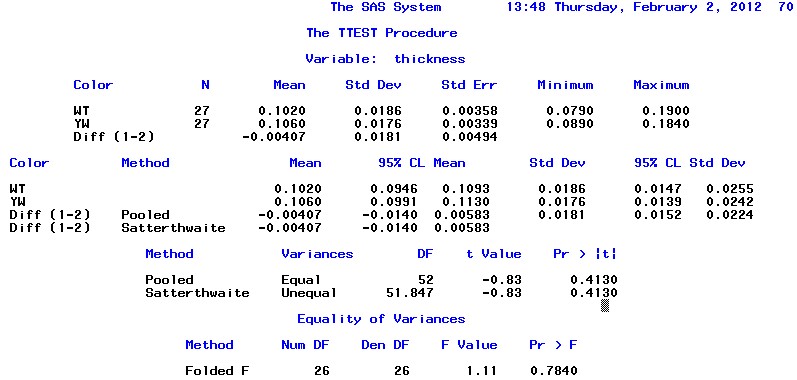
**4. Discuss the five steps of a 0.05-level paired-sample t-test to test the null hypothesis that there is no difference between the paper thicknesses. Show SAS or R output in your report. You may omit the outliers before performing the paired-sample t-test if you wish.**  


**Answer:**

1. State the null and alternative hypotheses
   1. H0: μd = 0 H1 : μd ≠ 0;
2. Compute the test statistic:
   1. t = = -2.35
3. Compute a 95% confidence interval I use the t-table with 26 degrees of freedom to show that [-0.00763, -0.00051] is a 95% confidence interval for t.
4. Determine to reject H0.
5. Let SAS compute the p-value: 0.0265 which is less than 0.05 thus we can be confident about our results.

**Notice: outliers are not omitted.**

**5. Discuss 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 the White and Yellow printer paper brands. Show your output and discuss what it means. Use the Reformat Example like you did for Project 1 to create a new dataset that is suitable for the independent two-sample t-test. Show SAS or R output in your report.**

****

**Anwser:**

1. State the null and alternative hypotheses

H0: μ1 =μ2 H1 : μ1 ≠μ2;

1. Compute the test statistic:

t= = = -0.83

1. Compute a 95% confidence interval I use the t-table with 52 degrees of freedom to show that [-0.014, 0.00583] is a 95% confidence interval for t.
2. Determine to reject H0.
3. Let SAS compute the p-value: -0.83, we are not so confident with our rejection.

**6. Is the paired sample or the independent two-sample t-test more appropriate to decide if the true thickness of a sheet of paper is different for Brand White or Brand Yellow? Justify your answer.  How do the p-values compare for the two tests.**

**Answer:** Those two p-values, 0.0256 for paired t-test and 0.7840 for independent, both note that we do not have enough confidence in the hypothesis that yellow and white papers have the same thickness, outliers considered. However, I prefer the independent t-test method since although they seem to be paired, they actually do not have a clear "pair" relationship between white and yellow.

**Part C. 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. (15 pts.) Explain some of the popular one and two-sample tests is to someone that is not familiar with statistics.  Discuss why you would want to use them and some of the important issues to consider.**

**Answer:**

One-Sample z-Test

A one-sample z-test is used to test whether a population parameter is significantly different from some hypothesized value.

Here is how to use the test.

* Define hypotheses. The table below shows three sets of null and alternative hypotheses. Each makes a statement about how the true population mean μ is related to some hypothesized value*M*. (In the table, the symbol ≠ means " not equal to ".)

|  |  |  |  |
| --- | --- | --- | --- |
| Set | Null hypothesis | Alternative hypothesis | Number of tails |
| 1 | μ = M | μ ≠ M | 2 |
| 2 | μ > M | μ < M | 1 |
| 3 | μ < M | μ > M | 1 |

* Specify significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level)equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Compute test statistic. The test statistic is a z-score (z) defined by the following equation.

z = (x - M ) / [ σ /sqrt(n) ]

where x is the observed sample mean, M is the hypothesized population mean (from the null hypothesis), and σ is the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=standard%20deviation)of the population.

* Compute P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a z-score, use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to assess the probability associated with the z-score.
* Evaluate null hypothesis. The evaluation involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

One-Sample t-Test

A one-sample t-test is used to test whether a population mean is significantly different from some hypothesized value.

Here is how to use the test.

* Define hypotheses. The table below shows three sets of null and alternative hypotheses. Each makes a statement about how the true population mean μ is related to some hypothesized value*M*. (In the table, the symbol ≠ means " not equal to ".)

|  |  |  |  |
| --- | --- | --- | --- |
| Set | Null hypothesis | Alternative hypothesis | Number of tails |
| 1 | μ = M | μ ≠ M | 2 |
| 2 | μ > M | μ < M | 1 |
| 3 | μ < M | μ > M | 1 |

* Specify significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level)equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Find degrees of freedom. The [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom)(DF) is:

DF = n - 1

where n is the number of observations in the sample.

* Compute test statistic. The test statistic is a t-score (t) defined by the following equation.

t = (x - M ) / [ s /sqrt(n) ]

where x is the observed sample mean, M is the hypothesized population mean (from the null hypothesis), and s is the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=standard%20deviation)of the sample.

* Compute P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t-score, use the t Distribution Calculator to assess the probability associated with the t-score, having the degrees of freedom computed above.
* Evaluate null hypothesis. The evaluation involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

### Two sample T-test

### Unpaired

The unpaired, or "independent samples" *t*-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared. For example, suppose we are evaluating the effect of a medical treatment, and we enroll 100 subjects into our study, then randomize 50 subjects to the treatment group and 50 subjects to the control group. In this case, we have two independent samples and would use the unpaired form of the *t*-test. The randomization is not essential here—if we contacted 100 people by phone and obtained each person's age and gender, and then used a two-sample *t*-test to see whether the mean ages differ by gender, this would also be an independent samples *t*-test, even though the data are observational.

### Paired

*Main article: Paired difference test*

Dependent samples (or "paired") *t-tests* typically consist of a sample of matched pairs of similar units, or one group of units that has been tested twice (a "repeated measures" *t*-test). A typical example of the repeated measures *t*-test would be where subjects are tested prior to a treatment, say for high blood pressure, and the same subjects are tested again after treatment with a blood-pressure lowering medication.

A dependent *t*-test based on a "matched-pairs sample" results from an unpaired sample that is subsequently used to form a paired sample, by using additional variables that were measured along with the variable of interest. The matching is carried out by identifying pairs of values consisting of one observation from each of the two samples, where the pair is similar in terms of other measured variables. This approach is often used in observational studies to reduce or eliminate the effects of confounding factors.

**2. (5 pts.) Why is it better to have a large sample than a small sample when analyzing data with methods that we discussed?**

**Answer:**

Larger sample size generates precise result when estimating unknown parameters. The larger the sample size, the more it normally distributes, according to the central limit theorem.

**3. (5 pts.) Explain what a one tailed z or t-test is and how is differs from a two-tailed test.**

**Answer:**

The two-tailed test is a statistical test used in inference, in which a given statistical hypothesis, H0 (the null hypothesis), will be rejected when the value of the test statistic is either sufficiently small or sufficiently large. This contrasts with a one-tailed test, in which only one of the rejection regions "sufficiently small" or "sufficiently large" is preselected according to the alternative hypothesis being selected, and the hypothesis is rejected only if the test statistic satisfies that criterion. Alternative names are one-sided and two-sided tests.

**4. (5 pts.) Statistical significance of a result is not always the same as practical importance. Explain what this means. Give an example if you can.**

**Answer:**

In statistics, a result is called "statistically significant" if it is unlikely to have occurred by chance.  Smaller levels of α increase confidence in the determination of significance, but run an increased risk of failing to reject a false null hypothesis (a Type II error, or "false negative determination"), and so have less statistical power. The selection of the level α thus inevitably involves a compromise between significance and power, and consequently between the Type I error and the Type II error.

**Notice: Main sources for referral are from google.com**