INVESTIGATING STATISTICAL VS. PRACTICAL SIGNIFICANCE OF THE KOLMOGOROV-SMIRNOV TWO-SAMPLE TEST USING POWER SIMULATIONS AND RESAMPLING PROCEDURES

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Lincoln Gary Larson

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

May 2018

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Investigating Statistical VS. Practical Significance of the Kolmogorov-Smirnov Two-Sample Test Using Power Simulations and Resampling Procedures

By

Lincoln Gary Larson

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Rhonda Magel
Chair

Curt Doetkott

Dr. David Hopkins

Approved:

May 15, 2018
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

This research examines the power of the Kolmogorov-Smirnov two-sample test. The motivation for this research is a large data set containing soil salinity values. One problem encountered was that the power of the Kolmogorov-Smirnov two-sample test became extremely high due to the large sample size. This extreme power resulted in statistically significant differences between two distributions when no practically significant difference was present. This research used resampling procedures to create simulated null distributions for the test statistic. These null distributions were used to obtain power approximations for the Kolmogorov-Smirnov tests under differing effect sizes. The research shows that the power of the Kolmogorov-Smirnov test can become very large in cases of large sample sizes.

ACKNOWLEDGEMENTS

I would like to thank Curt Doetkott for advising me while writing this thesis and guiding me in my learning, during my time at North Dakota State University. I would also like to thank Dr. Magel for her guidance in preparing this thesis and completing my research. I would also like to thank Dr. Magel for her contributions that expanded my knowledge in the areas of mathematical and non-parametric statistics. Thanks, are also in order for Dr. Hopkins, who was on my thesis committee and explained how the original salinity data was collected.

I would also like to thank the many other people that helped me expand my knowledge including my dad, mom, and siblings. Thanks, are in order for my grandparents and dad for picking up a large portion of my responsibilities on the farm, so I could move to Fargo and continue my education. I would also like to thank both sets of my grandparents for encouraging me to continue my education. I also wish to thank Dr. Dan Johnson for his contributions to my understanding of mathematics.

DEDICATION

This thesis is dedicated primarily to Jesus Christ my savior. This thesis is also dedicated to my mother, who worked to educate me from the time I was born and homeschooled me for many years. I would also like to dedicate this thesis to my dad, grandparents, and the rest of my family. Thank you all for your encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

GPS ....Global Positioning System

EC ....Electrical Conductivity

KS ....Kolmogorov-Smirnov

SAS ....Statistical Analysis System

ES...Effect Size (defined as $\Delta\mu/\sigma$)

N...Sample Size

LIST OF SYMBOLS

α ....(alpha) A Greek letter used to represent the significance level or type one error.

μ ....(mu) A Greek letter used to represent the population mean.

σ ....(sigma) A Greek letter used to represent the population standard deviation.

Δ ....(delta) A Greek letter used to represent change in a value following the delta.

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

1. INTRODUCTION

With modern electronic data collection methods, it is common to have very large data sets. It is important to understand how large sample sizes can affect the results of different analysis methods. If a researcher fails to understand how large sample sizes affect the results of a specific method, the results could be confusing or unmeaningful in a practical sense. This thesis will specifically focus on the two-sample Kolmogorov-Smirnov test and how sample size affects the test's power.

It may be desirable to test if two population distributions differ significantly. One way to approach this type of problem is to use the Kolmogorov-Smirnov two-sample test. This test allows one to compare the empirical distributions of the two populations. The hypothesis statements that may be used with the Kolmogorov-Smirnov two-sample test are as follows:

$$Ho: F_1(x) = F_2(x) \quad \text{For all } x$$
$$Ha: F_1(x) \neq F_2(x) \quad \text{For at least one } x$$

In this case, $F_1(x)$ represents the cumulative distribution function of the first population and $F_2(x)$ represents the cumulative distribution function of the second population.

The test statistic for the two-sided test is the absolute value of the maximum distance between the empirical distributions of the two populations.

$$\mathrm{D} = \max \left| S_1 - S_2 \right|$$

S_1 represents the empirical distribution for population one and S_2 represents the empirical distribution of population two. D is the maximum absolute value of the distances between the two empirical distributions.

The null hypothesis is rejected if the test statistic is larger than the, 1 - α, quantile of the null distribution (Daniel, 1990). α represents the desired level of significance. In this thesis α of 0.05 will be the only level of significance considered. Rejecting the null hypothesis indicates that there is a statistically significant difference between the two empirical distributions. If the null hypothesis is not rejected, it indicates that we do not have evidence to say the two populations differ significantly.

The motivational data set used in this thesis came from soil salinity research, in North Dakota. Data was collected by Hopkins and Steele using a veris machine, which was attached to the rear of a pickup (Hopkins & Steele, 2011). This machine collects soil electrical conductivities, which is "a measurement of how much electrical current soil can conduct" (Veris Technologies, Inc., 2014). The veris machine collects soil conductivity data by running electrical current through coulters in the soil. The electrical conductivity values are then recorded electronically at two different levels, shallow and deep. Shallow represents the depth from 0 to 12 inches and deep represents the depth from 0 to 36 inches. In general, the higher levels of soil conductivity represent higher levels of salinity in the soil. Other variables recorded with the veris include, GPS latitude and longitude, elevation, and a sample identification number. This thesis focuses mainly on the electrical conductivity variables, deep and shallow. The focus is mainly on a few selected sites. These sites are analyzed to get a better understanding of how the sample size and absolute effect size affects the power of the Kolmogorov-Smirnov two-sample test.

Hopkins and Steele (2011) collected data from sites once in the spring and once in the fall from the years 2005 to 2009, when access to the land was available. Because of access issues researchers were unable to obtain data for every season. Data was collected from a total of eight sites. The goal of the original analysis was to determine if there was a difference in the site's soil salinity levels over time. The Kolmogorov-Smirnov test was used to determine if there was a difference within sites over time. When the original tests were performed the null hypothesis was always rejected, which indicates a statistically significant difference in the salinity levels was present.

This thesis investigates why the null hypothesis may have been rejected so frequently. It is suspected that the large sample sizes are responsible for a substantial increase in the power of the test. The goal is to determine how the sampling effort affects the power of the test. The method used to determine if this is in fact the case, will be simulations using SAS software. The methodology of these simulations is explained in greater detail in Chapter 3 of this thesis.

2. LITERATURE REVIEW

Extensive research has been done on the power of the Kolmogorov-Smirnov two-sample test. This section will highlight some of the research on the power of the Kolmogorov-Smirnov test that has previously been published. There has been a large amount of research done on the power of the KS test compared to the power of other similar tests. One of the most recent research papers focused on using Monte Carlo simulations to find the power of the Kolmogorov-Smirnov test for different types of distributions and different distribution parameters (Boyerinas , 2016).

Boyerina's simulation research focused on normal, lognormal, and exponential distributions. He looked at the power of the KS test and the Anderson-Darling test. He shows that the Kolmogorov-Smirnov test is more sensitive to differences near the center of the empirical distribution functions than the Anderson-Darling test. However, the Anderson-Darling test is more sensitive in the tails of the empirical distribution functions. It is necessary to simulate data from the distributions, under different distribution parameters, to get an accurate estimation of the power of a test.

The results of Boyerina's Monte Carlo simulations showed that the power of the two-sample KS test varied significantly depending on the mean and variance parameters. It was shown that the Anderson-Darling test, in general, has a slightly higher power than the two-sample KS test under the same distribution parameters. The power of the KS test was found to change significantly depending on the distribution parameters of mean and variance. The power always increases with larger sample sizes, as can be expected. The absolute effect size affects the power of the KS test significantly.

The power of a statistical test is very important. If the power is too low the inferences made from the test may not be correct (Massey Jr., 2012). It is important to analyze not only the statistical significance of a test, but also its practical significance. Ellis (2010) notes that, most studies are completed, and conclusions are drawn without even considering the power of their inferential tests. If the power of a test is too low, for the sampling effort in a study, a practically significant result may not be detected when it should be.

Absolute effect size refers to the difference of the sample means divided by the sample standard deviation. If the absolute effect size is large enough the results of a test will be practically significant (Ellis & Steyn, 2013). Smaller absolute effect sizes may not necessarily be practically significant. If the power of a test is high the test may find small absolute effect differences to be statistically significant. These statistically significant differences may not necessarily be significant is a practical sense.

“Effect size can be considered an index of the degree to which the findings have practical significance in the population study” (Hojat & Xu, 2004). Actual effect size for the estimate of mean differences is the ratio of the differences in the means over the standard deviation of the control group (Nakagawa & Cuthill, 2007). Hojat and Xu note that many research journals have started to recommend or even require authors to submit effect size estimates.

It is notable that the Kolmogorov-Smirnov test does not result in continuous test statistics for two samples of equal size (Daniel, 1990). If the sample sizes are equal the test will result in discrete test statistics (Boyerinas , 2016). It desirable to use continuous test statistics for the null distribution, so it is important to use unequal sample sizes in simulation studies of the two-sample KS test.

Statistical significance represents an improbable result, but practical significance represents a difference that is meaningful in a real-world sense (Ellis P. D., 2010). Ellis notes in his book that few researchers “distinguish between the statistical and practical significance of their results." Practical significance refers to the usefulness of a result in a real world situation, but statistical significance is related to the ability to detect differences between two samples.

Research has been done to compare the Kolmogorov-Smirnov test power to the power of other tests. There are also many papers on the efficiency of the KS test (Klotz, 2012) (Capon, 1965) (Ramachandramurty, 1966) (Yu, 1971). Simulations have been used to determine the power of the test under different distribution parameters. It is difficult to find any publications on the effect of large sample sizes for the KS tests. This thesis examines the extreme power caused by large sample sizes as well as the importance of practical vs statistical significance.

3. METHODOLOGY

3.1. Data Description

The data used in this thesis was briefly discussed in the introduction of this paper. This section gives a more detailed description of the data used in this thesis. The data from the motivational data set came from a research study conducted by Hopkins and Steele (2011). They collected soil electrical conductivity data using a veris machine. This machine was attached to the back of a pickup. The pickup followed a GPS system to cross a field in a back and forth pattern with equal distances between each parallel pass. As the machine moves it pulls coulters through the soil. These coulters periodically record the soil's electrical conductivity at two different depths. The shallow depth is defined from 0 to 12 inches and deep refers to the depth from 0 to 36 inches. This data was recorded once in the fall and once in the spring for each field, when conditions allowed.

Each measurement that is recorded also has an ID number, latitude, longitude, and elevation value. The electrical conductivity values are measured as apparent electrical conductivity. The values of EC recorded from the selected sites ranged from 0.2 to 626.3. The smallest number of observations from the selected sites was 2,917 and the largest number of observations was 15,906. Most of the electrical conductivity data has a substantial amount of variation. The data sets also tend to be very right skewed.

3.2. Site Selection

The sites chosen to analyze were determined based on certain criteria. It was decided that sites with different distribution parameters, mean and variance, and with different distribution types would be selected. Site 12 fall 2007 deep data was chosen because the number of observations is high, 15,906. The deep EC values from this data were used for the simulations. Site 35 spring 2008 was chosen because the values of deep EC had a somewhat bimodal distribution. Site 35 spring 2007 shallow data was chosen because it has a large variation compared to the other chosen sites. This data is also very right skewed. The last site chosen for the simulations was the shallow values for site 20 fall 2008. This site was chosen because it was very right skewed with a very small variance in comparison to the other sites. Appendix C contains more information on the selected sites and their distributions including descriptive statistics.

3.3. Simulation Procedure

The procedures used in this thesis can be broken down into three major steps. The three steps involved are:

- Simulations to calculate approximate null distributions
- Simulating samples and calculating test statistics
- Calculation of rejection rates for samples, using the approximated null distribution

The following section of the thesis explains in detail how each step above can be implemented in order to obtain useful results.

3.3.1 Calculation of Approximate Null Distributions

The null distributions of the test statistic are simulated for each sample size used in this thesis. For every combination of sample sizes, 100,000 test statistic values are calculated when the condition of the null hypothesis is true. The 95th percentile of each set of 100,000 test statistic values is used as the critical value for tests performed at that site for the particular sample size combination. The distribution of the test statistic is only continuous when the sample sizes comparing two distributions are not equal (Boyerinas , 2016). Because of this, the combinations of sample sizes used in this thesis are (25, 26), (50, 51), (100, 101), (200, 201), (500, 501), (1000, 1001), and (2000, 2001). It is assumed that this slight difference in sample sizes will have minimal effect on accuracy of the power calculations, as unequal sample sizes are actually an assumption of the two-sample KS test (Daniel, 1990).

The basic form of the SAS code used for the reference distributions is found in Appendix D.1. This code uses a seed of 0 to randomly select observations from the desired data set. Setting the seed to zero tells SAS to use the clock value as the random seed. Observations are randomly selected for sample 1 and sample 2. Once the samples are obtained, the test statistics are calculated for the two-sample Kolmogorov-Smirnov tests using proc npar1way in SAS (SAS Institute Inc., 2010). There are 100,000 test statistics calculated for each sample size at each of the chosen site's distribution. This collection of test statistics is the approximated null distribution for its corresponding site and sample size. The percentiles of these approximate null distributions are then found using the code in Appendix D.2. The 95th percentile value of the null distribution is saved for use as a critical value in future steps of this thesis.

For an example we will look at the sample size of 25 and how this null distribution was calculated. Two random samples are taken from the original data set. Sample one has a sample size of 26 and sample two has a sample size of 25. These two samples are then run through the npar1way procedure in SAS. This procedure calculates the test statistic for the two-sample KS test. This test statistic is the maximum difference between the empirical distributions of the two samples. This process of taking two samples and calculating the test statistic is done 100,000 times. This results in 100,000 test statistics for the sample size of 25. The 95th percentile of the 100,000 test statistics is then calculated. This 95th percentile value is the critical value of the test. This process is repeated for each of the sample sizes we wish to approximate power for.

The test statistic that is used in this thesis is called the KSa. The KSa is calculated from the empirical distributions. The empirical distribution calculation is shown below.

$$\mathrm{F}(x) = \frac{1}{n}\sum_{i}(n_i F_i(x))$$

"Where n_i is the number of observations in the i^{th} class level and n is the total number of observations" (SAS Institute Inc., 2010). The formula for the KS test statistic used in SAS follows.

$$KS = \max_{j}\sqrt{\frac{1}{n}\sum_{i}(n_i F_i(x_j) - F(x_j))^2}$$

The KSa value is what is obtained from the simulations in this thesis. The KSa is just the square root of n times the KS value. This KSa test statistic is simply an asymptotic version of the KS test statistic.

$$KSa = KS \times \sqrt{n} = \sqrt{n} \times \max_{j}\sqrt{\frac{1}{n}\sum_{i}(n_i F_i(x_j) - F(x_j))^2}$$

3.3.2. Simulating Samples and Calculating Test Statistics

The samples are created using a bootstrapping type of approach. This method involves sampling with replacement. The code used to create these samples is provided in Appendix D. The same sample sizes were used as those used in the approximate null distribution creation, with $n_1 = n_2 + 1$. Sample sizes for n_2 were 25, 50, 100, 200, 500, 1000, and 2000. These sample size values are the n_2 values. 10,000 samples were taken for each combination of absolute effect size and sample size at each site. These samples are then run through proc npar1way in SAS to calculate their KSa test statistics. The calculated test statistics are saved for use in the last step of the procedure.

These test statistics are calculated for different absolute effect sizes. Absolute effect size in this thesis is defined below.

$$\mathrm{ES} = \frac{\Delta\mu}{\sigma}$$

Where ES denotes absolute effect size and delta mu is the difference between the means. As shown in the above formula, absolute effect sizes are related to the original sample's mean and standard deviation. The absolute effect sizes chosen for this thesis are 0, 0.1, 0.2, 0.5, and 0.8. These absolute effect sizes are simulated for each of the sample sizes, 25, 50, 100, 200, 500, 1000, and 2000, for each site.

The term "actual effect" will be used in this thesis to describe the magnitude of the difference in the means. This actual effect is simply the change in means. Actual effect will be denoted as $\Delta\mu$.

To see how actual effect size affects the power, data will be simulated for site 35 spring 2007's shallow distribution using the actual effects from site 20. These actual effects are the $\Delta\mu$ values used in the simulations of site 20. This simulation using actual effects will be useful in understanding how the actual effect affects the power of the KS two-sample test.

Following the above steps will provide the simulation data needed to calculate the power of the two-sample KS test. It is essential to combine the samples by effect size using the code in Appendix D.4 to plot all of the data on one plot. In the next step these test statistics are compared to the 95th percentile of the corresponding simulated null distribution to calculate the power of the KS test. The power will be calculated for each combination of sample size and effect size.

3.3.3. Calculation of Rejection Rates

The power will be found by finding the proportion of rejections of Ho when Ha is true. This proportion is found by comparing the sample KSa test statistics to the 95^{th} percentile of the KSa test statistics from the approximated null distribution, for the corresponding sample size. This proportion is also calculated for the absolute effect size of 0, which represents the type one error. This type one error is expected to be approximately 5%. The estimated type one errors from these simulations are provided in Appendix A.

3.4. Power Plot Creation

The power of the KS tests is plotted using the gplot procedure in SAS (SAS Institute Inc., 2016). These plots are made for each site separately. The x-axis is the sample size and the y-axis is the power. These power plots contain four different lines. Each line represents one of the effect sizes. These power plots are provided in the results section of this thesis.

The plots for the type one error are created using the same method. Type one error should be very close to a 5% rejection rate. It is important that the type one errors be consistent so that results from different simulations will be comparable. These type one error plots should result in a horizontal line at the 5% rejection rate. The code used to create these plots is in Appendix D.5. The critical values calculated from the null distribution are hard coded into this code. The power plots are shown in Chapter four of this thesis.

4. RESULTS

4.1. Site 35 Spring 2008 Deep Distribution

Site 35's spring 2008 deep distribution was chosen as one of the data sets used to simulate power for. This distribution was chosen because it is somewhat bimodal with a large variance. The histogram of this distribution can be seen in Figure 1. The mean for this distribution is 103.2 with a standard deviation of 77.79. The original number of observations for the site 35 distribution was 3077.

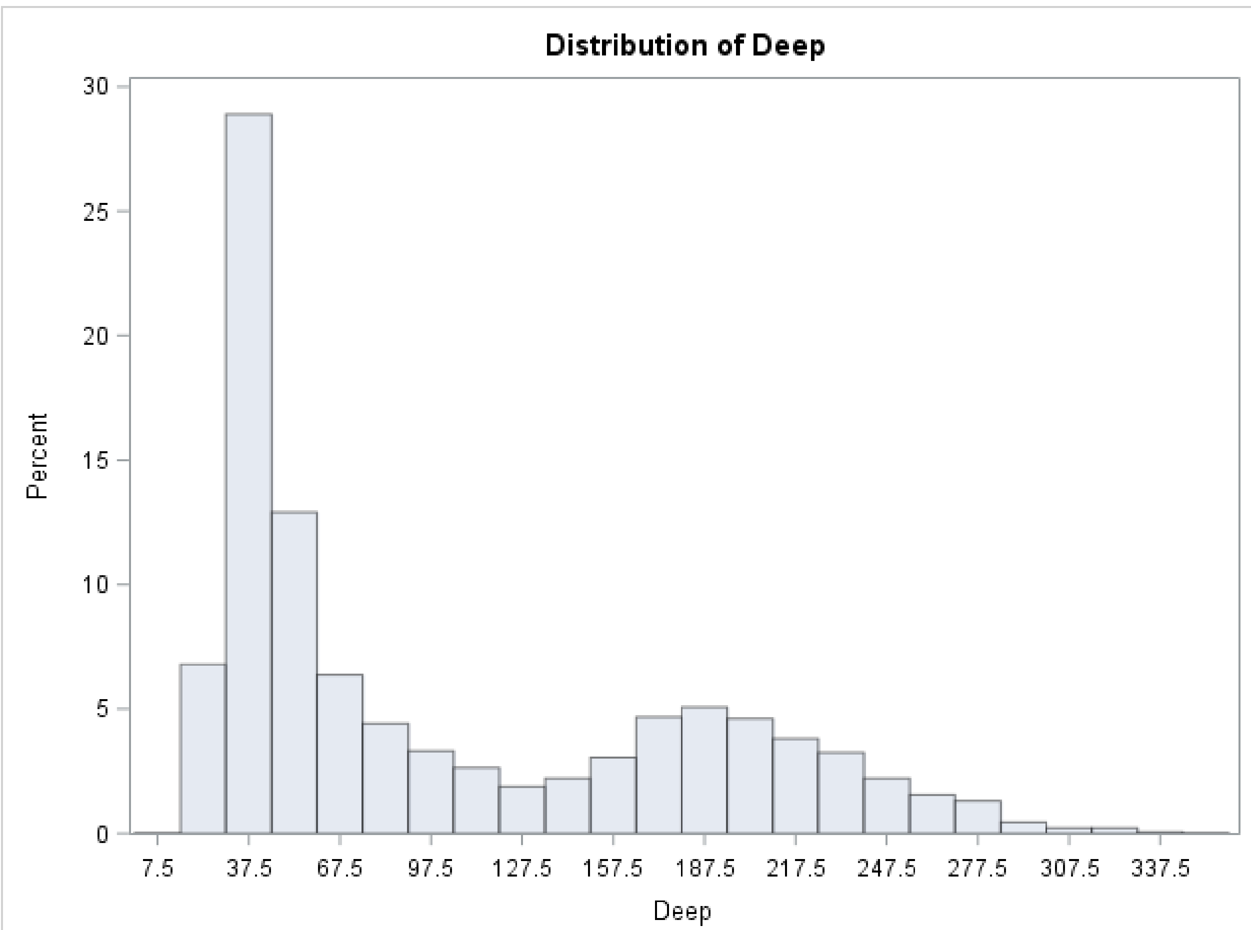

Figure 1. Site 35 Spring 2008 Deep Distribution

The procedures in chapter 3 were followed to create the null distribution for each of the chosen sample sizes 25, 50, 100, 200, 500, 1000, and 2000. The percentiles of the KSa test statistics were found using the code in Appendix D. Data was then resampled for each of the absolute effect sizes 0, 0.1, 0.2, 0.5, and 0.8. These effect sizes were implemented by adding each effect size times sample standard deviation of 77.79 to each value simulated in sample 2. The $\Delta\mu$ for each effect size is 0, 7.78, 15.56, 38.89, and 62.23 respectively. The KSa values for each of these simulations was compared to the 95th percentile of the KSa values in the approximated null distribution, with the corresponding sample size. The power was calculated as the percent of rejections when the absolute effect size was not equal to 0. This percent represents the correct rejection rate. The power plot that was created is shown in Figure 2.

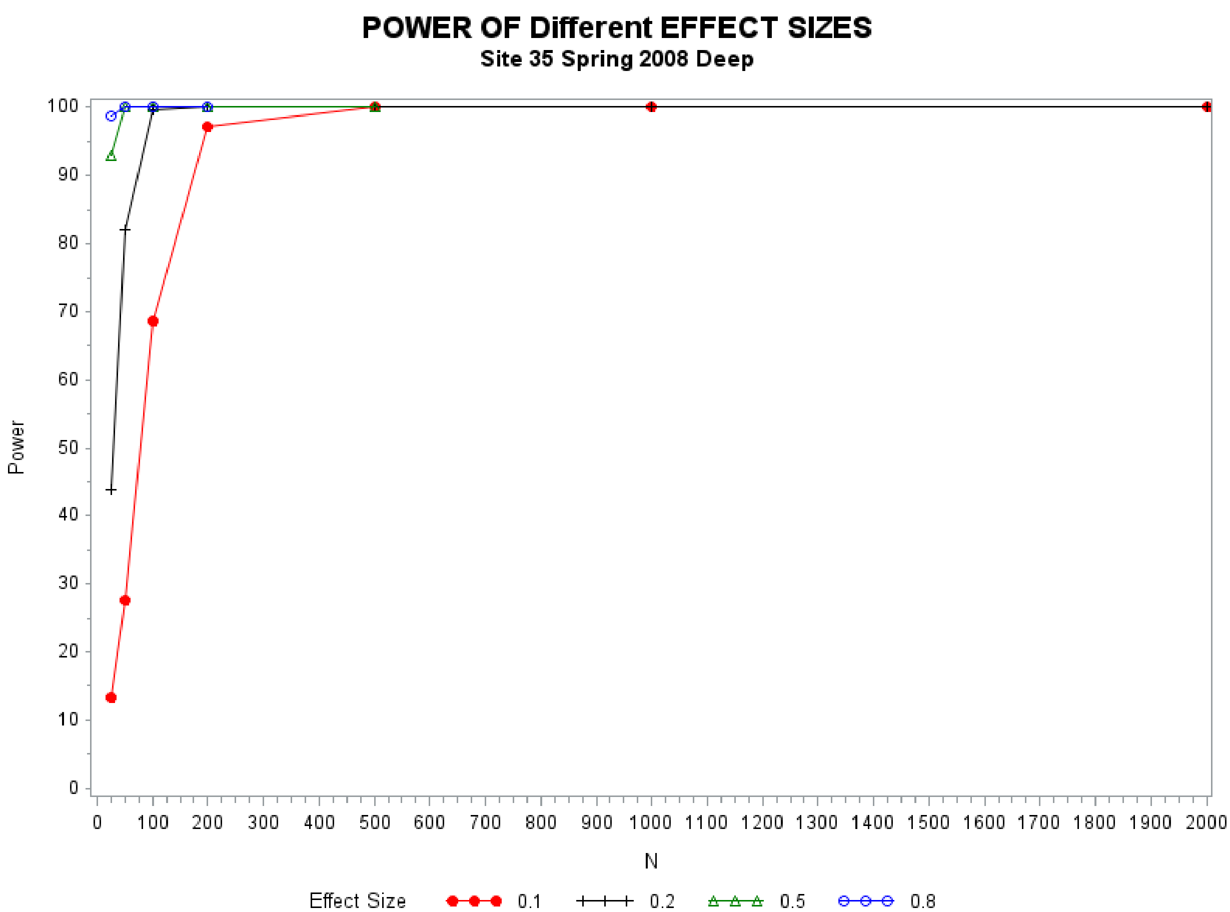

Figure 2. Site 35 Spring 2008 Deep Power Plot

The power plot clearly shows that the power increases with sample size for all absolute effect sizes. Larger absolute effect sizes indicate a larger level of practical significance. The smaller the effect size the smaller the power is, for a given sample size. A power of 100% is reached at fairly small sample sizes. Even the lowest absolute effect size of 0.1 yields about 100 percent power at sample sizes of only 500. With a larger absolute effect size, such as 0.8, the power of the test is high even at the smallest sample size of 25. These results show that the power of the KS test can be very high even with small sample sizes. This was the result we expected based on the results of the original KS tests done on the salinity data.

4.2. Site 35 Spring 2007 Shallow Distribution

This distribution at site 35 was chosen because it has a very large variance compared to the other distributions. This distribution is right skewed which can be seen in Figure 3. The mean for this distribution is 110.77 and the standard deviation is 114.4. Since this site has such a large variance the difference in the means will be greater. This results in larger actual effect sizes, which will cause the power of the test to increase.

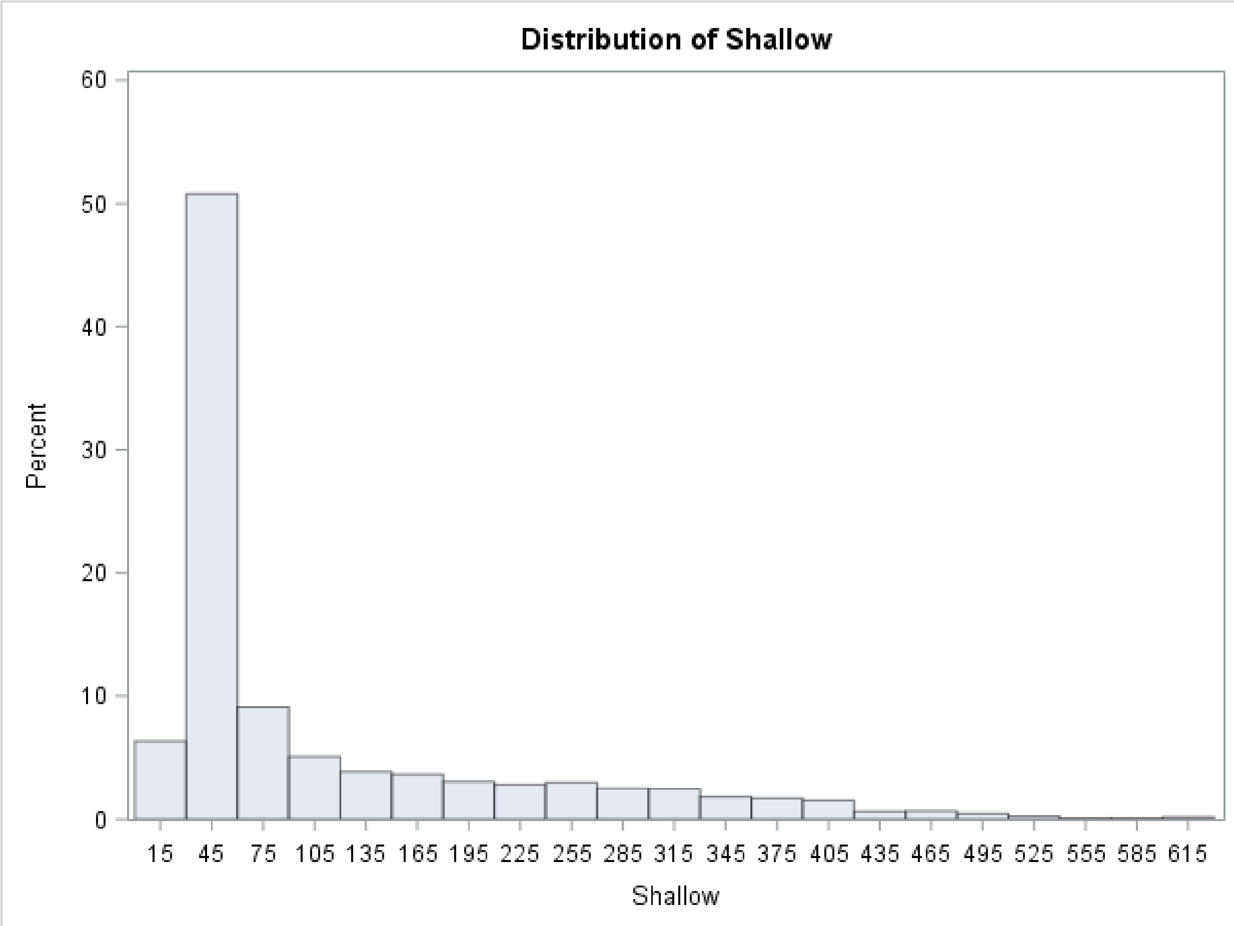

Figure 3. Site 35 Spring 2007 Shallow Distribution

The actual effect values added were 0, 11.44, 22.88, 57.20, and 91.52. These correspond to the absolute effect sizes of 0, 0.1, 0.2, 0.5, and 0.8 respectively. The power plot that was created for this distribution is presented in Figure 4.

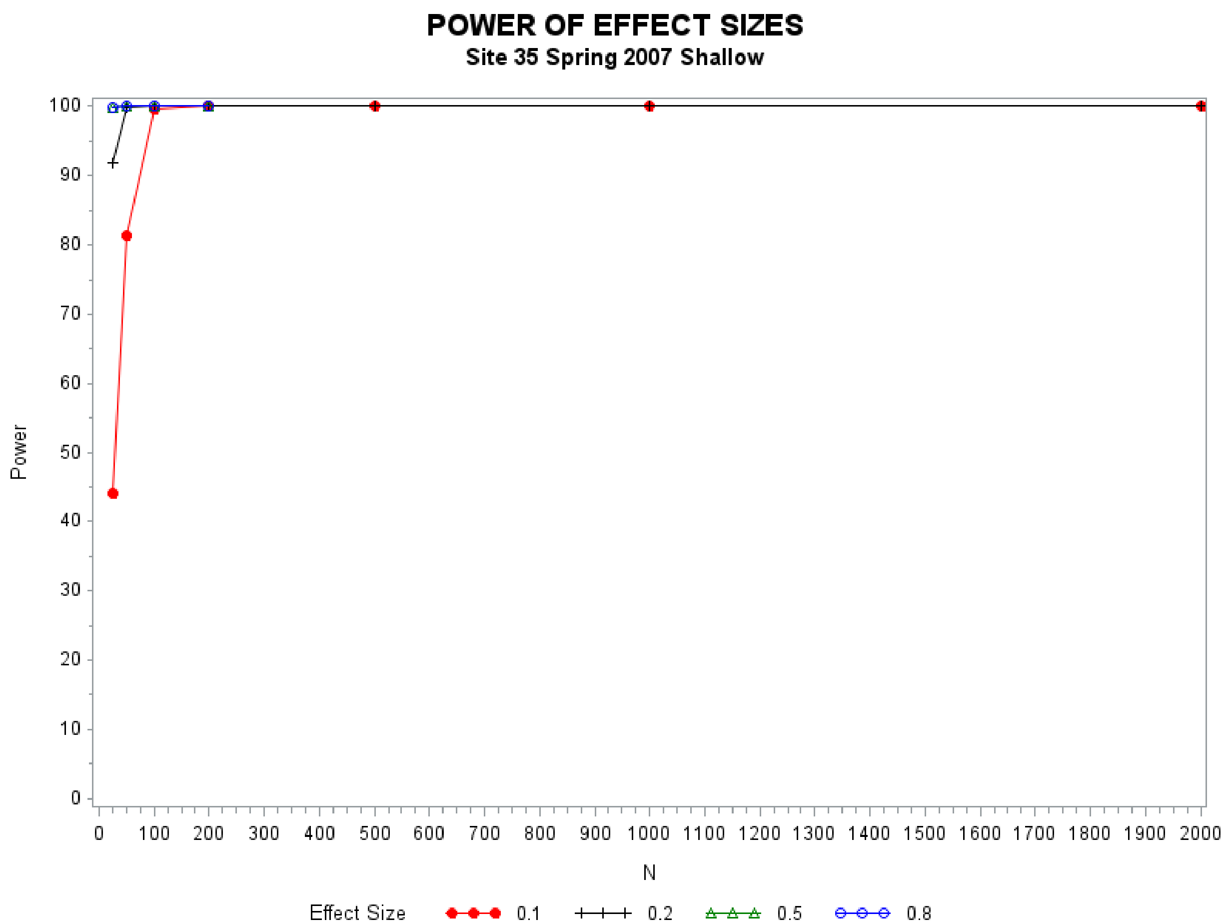

Figure 4. Site 35 Spring 2007 Shallow Power Plot

Looking at the power plot for this distribution it is easy to see that there is very high power for small sample sizes. Even with the smallest absolute effect size of 0.1 the power is over 80 percent with a sample size of only 50. For larger absolute effect sizes the power is almost 100 percent at the smallest sample size of 25. If the KS test was used for a distribution like this with large sample sizes it may result in practically insignificant rejections of Ho. This plot helps to explain how the large power could cause practically insignificant rejections.

4.3. Site 20 Fall 2008 Shallow Distribution

The fall 2008 shallow distribution was chosen for site 20 in the original data. This distribution was chosen because it has a lower variance than most of the other distributions.

Since it has lower variance, the difference between the means for the two samples will be smaller, for a given absolute effect size. The mean of this distribution is 28.8 and the standard deviation is 9.33. This histogram for this distribution is provided in Figure 5.

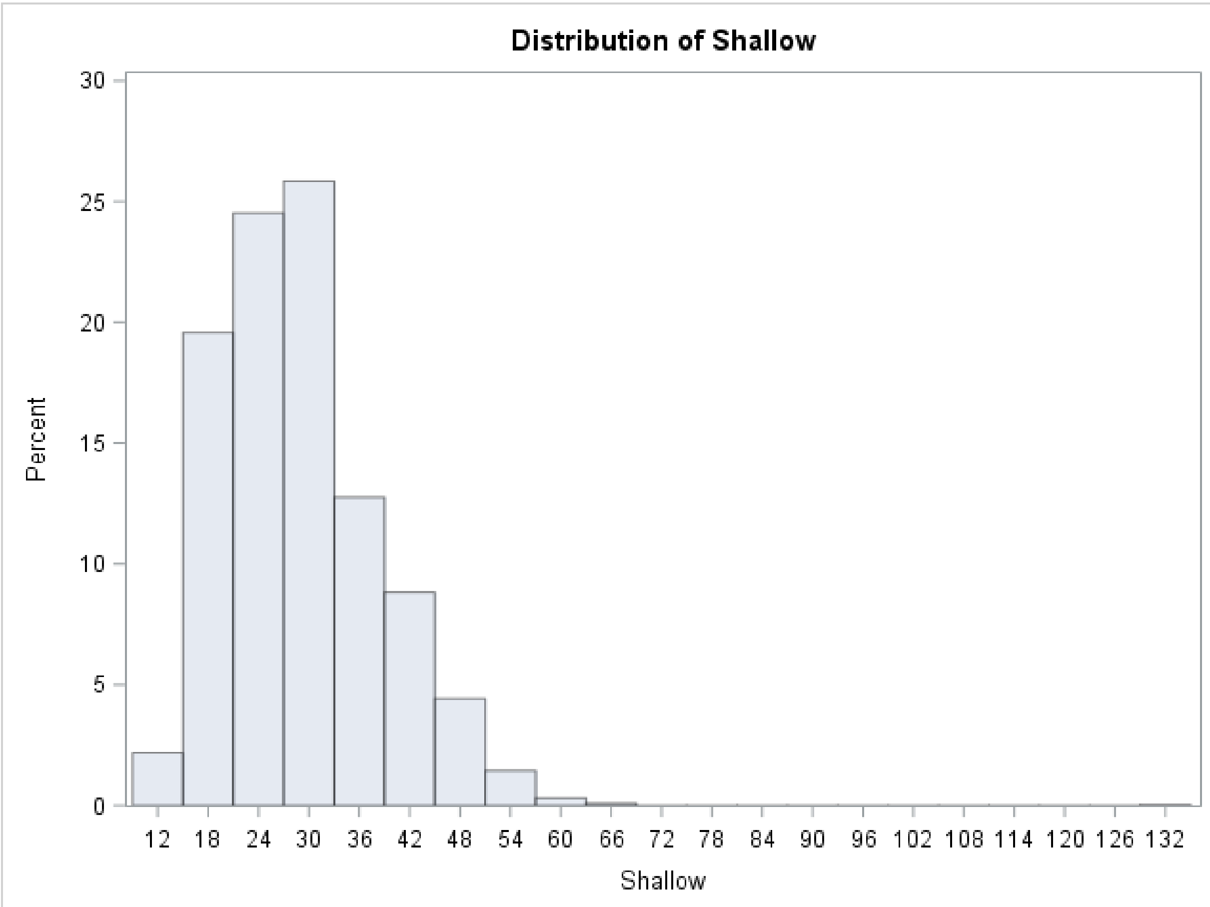

Figure 5. Site 20 Spring 2008 Shallow Distribution

The values that were added to the sample two to create the absolute effect sizes were, 0.93, 1.87, 4.67, 7.47. Using the methods that are outlined in the procedure section, the power plot in Figure 6 was created.

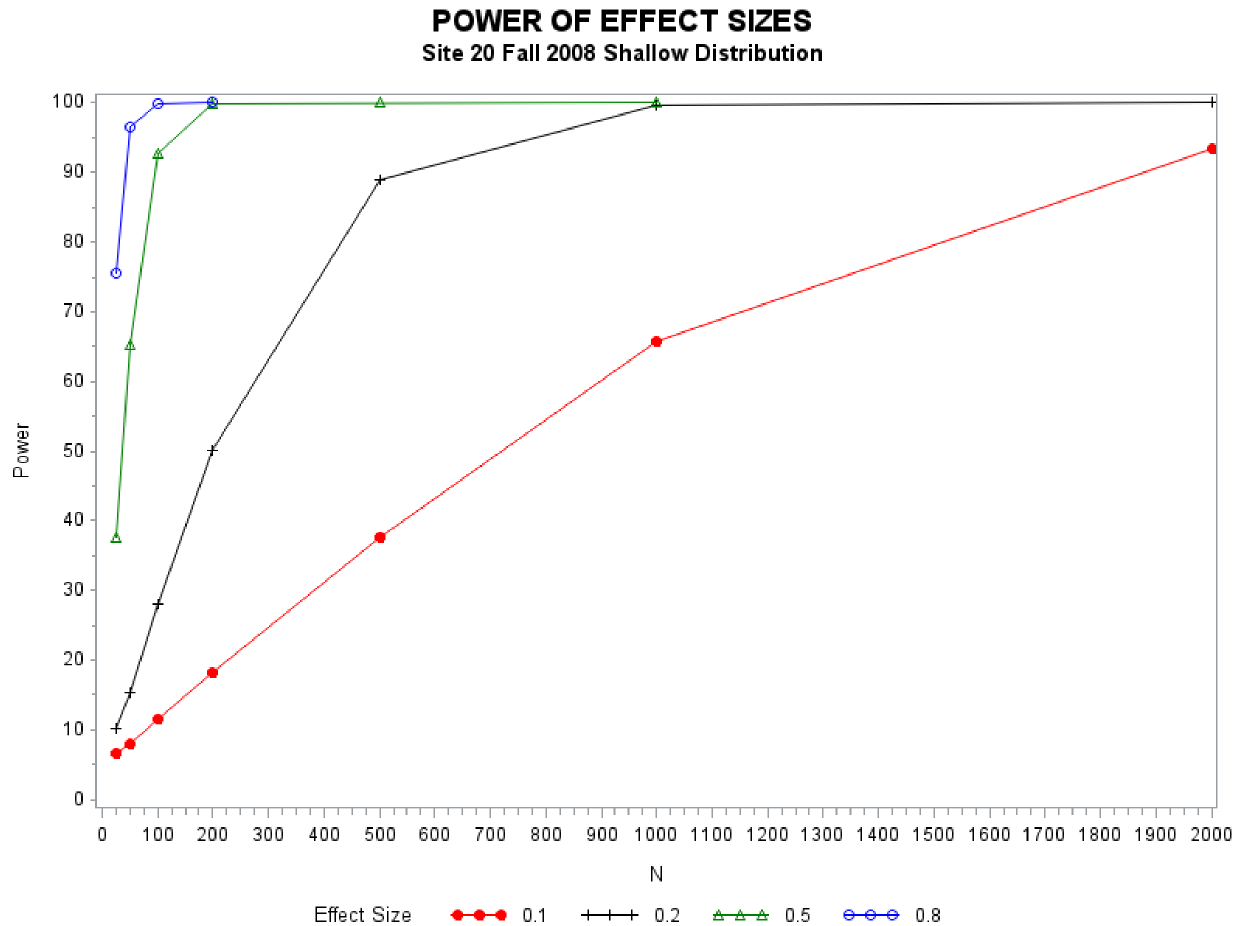

Figure 6. Site 20 Fall 2008 Shallow Power Plot

The power plot for site 20 has much lower power for the same absolute effect sizes when compared to the previous distributions. With an absolute effect size of 0.1 the power of the test is not even 70% for the sample size of 1000, but the power still increases to over 90% at sample sizes of 2000. For larger absolute effect sizes the power is much higher. With an absolute effect size of 0.8 it is only necessary to have sample sizes of 50 to get a power greater then 90%. It becomes very apparent that the effect magnitude, actual effect, one desires to detect must be known before choosing the most appropriate sampling effort for this test.

4.4. Site 12 Fall 2007 Deep Distribution

This site is right skewed with a mean of 52.6 and a standard deviation of 37.86. The number of original observations for this distribution was 15,906. The histogram for this distribution is provided in Figure 7. The values that were added to the sample two values, to create the effect sizes were, 3.79, 7.57, 18.93, and 30.29. This means that the change in the mean, $\Delta\mu$, is not as large as some distributions at other sites. The power plot for this distribution can be seen in Figure 8.

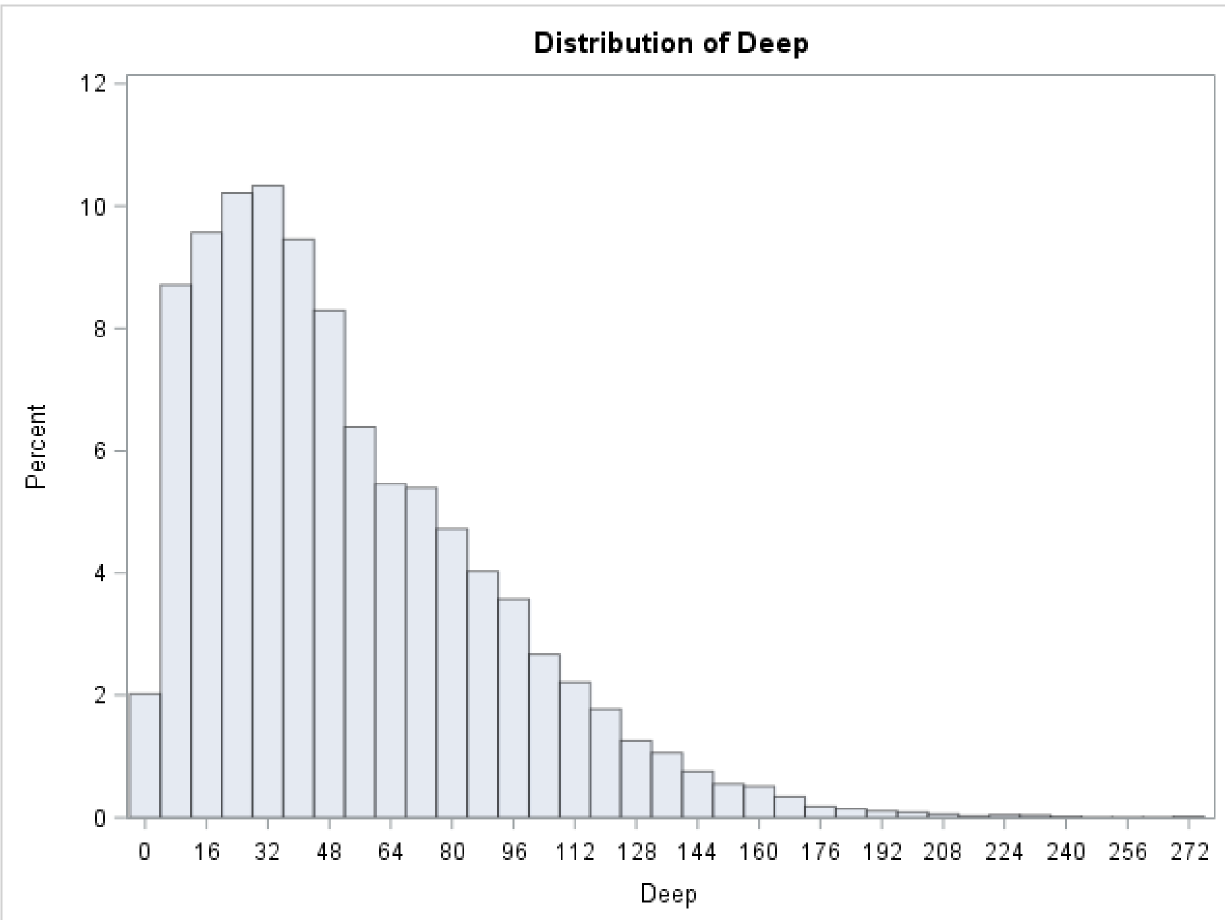

Figure 7. Site 12 Fall 2007 Deep Distribution

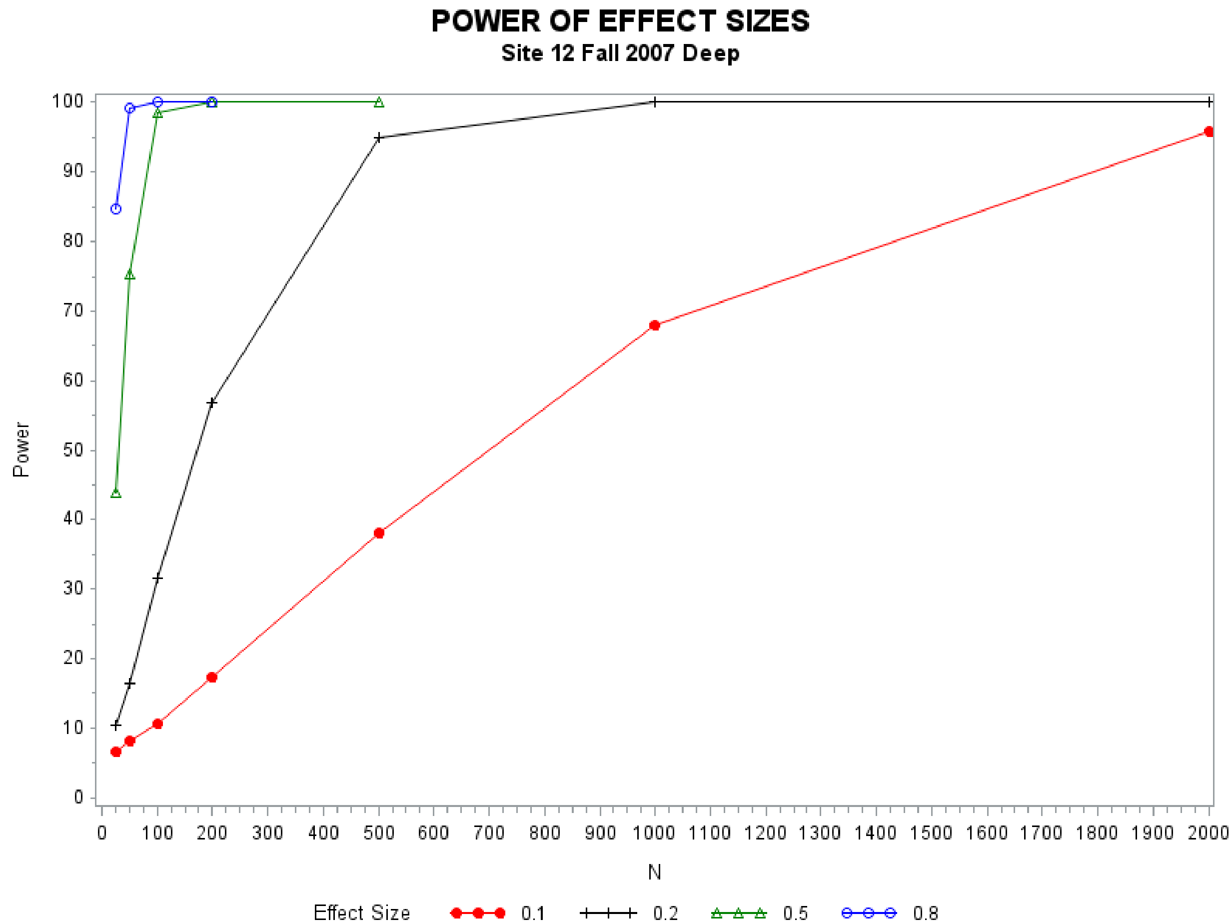

Figure 8. Site 12 Fall 2007 Deep Power Plot

The power plot shows that the power is quite high for large absolute effect sizes, like 0.8, even with relatively small sample sizes. The absolute effect size of 0.1 does not have a high power until sample sizes reach approximately 1,500. The next result is the power plot for actual effect size. With this result it is easier to see how actual effect changes the power of the KS two-sample test.

4.5. Site 35 Spring 2007 Shallow Distribution with Actual Effects

The actual effect refers to the actual magnitude of the difference of means, $\Delta\mu$. It was desired to see how this actual effect changes the power of the two-sample KS test. The data used for this simulation is that from site 35's spring 2007 shallow distribution. In this case, instead of adding the absolute effect size, which was done in section 4.2, we add the $\Delta\mu$ from the site 20 distribution, from section 4.3. These $\Delta\mu$ values are 0.93, 1.87, 4.67, and 7.47. When these actual effects are used it creates the power plot in Figure 9.

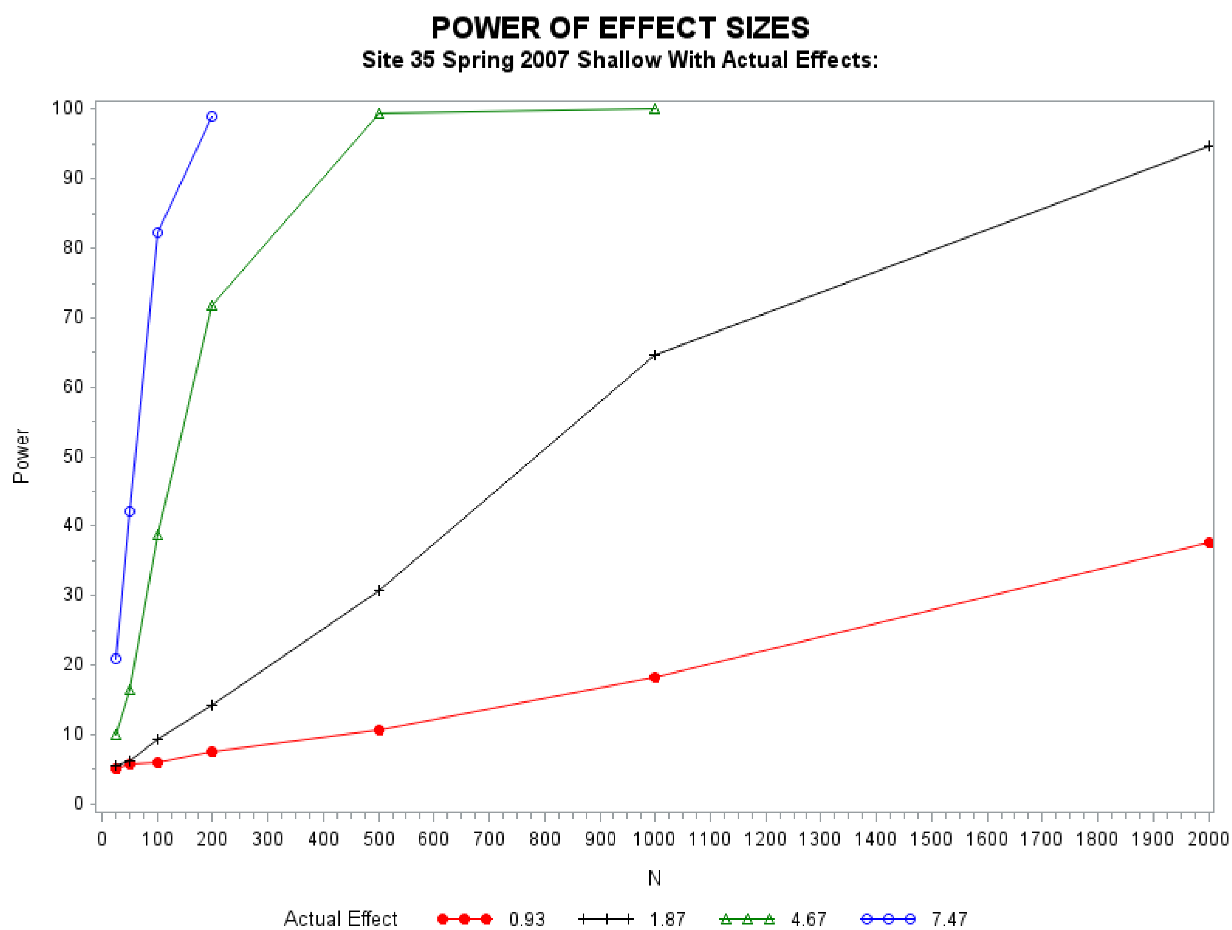

Figure 9. Site 35 Spring 2007 Shallow Power Plot for Actual Effects

When comparing this power plot to the power plot created using absolute effects we see that this plot has lower power. This is the expected result, because the values of $\Delta\mu$ for site 20 are smaller than the differences of the means that were used when finding the power for the absolute effect size.

Comparing this power plot to the one created for absolute effect sizes for site 20, Figure 6, we see that there is lower power for this plot of actual effects. This result makes sense because site 35's spring 2007 distribution has a much larger variance than that of site 20. It can be expected that the power of a test will decrease as the variance of the population's distribution increases. This result helps explain how the power of the two-sample KS test reacts to actual differences between the means.

5. CONCLUSION

After completing the simulation procedures, outlined in the methodology section, and creating the power plots, displayed in the results section, some conclusions can be made. It is obvious that the mean and variance of the distributions being tested has an extreme effect on the power of the test in terms of absolute effect size. It is important, for researchers, to have an understanding of how these parameters of mean and variance may affect the power of the two-sample Kolmogorov-Smirnov test. Without this understanding a researcher may make impractical claims about the differences between two populations.

To better explain how this understanding is important, an example using the site 12 fall 2008 shallow distribution is given. If a researcher desires to compare two distributions, say fall 2008 to fall 2007, the number of observations for each distribution would be more than 1,000. Let us also assume that the electrical conductivity values are only accurate within 5 units. If the researcher then uses a computer program to run the two-sample KS test, with each population having more than 1,000 data points, the result will likely be to reject the null hypothesis even for a very small difference. This statistically significant difference could be much smaller than the accuracy of the veris machine's electrical conductivity values. This could be a problem because the null hypothesis will essentially be rejected when no practically significant difference is present. The only real difference being detected could be the natural variation in the data. This problem of large sample size has the possibility of making the statistical inference useless in a practical sense. Refer to Figure 6 from site 12, to see how small differences in $\Delta\mu$ will cause a rejection at larger sample sizes.

The absolute effect size of 0.1 in the power plot in Figure 6 represents an actual effect of approximately 3.8 units of electrical conductivity. The ability to detect this difference of 3.8 units is over 95% with sample sizes of only 2000. It is easy to see that the power would essentially reach 100% with the sampling effort in our example with over 1,000 observations. This example clearly shows that it is possible for a two-sample KS test, with large sample sizes, to lead to rejections that are not practically useful. If the sample sizes were even larger this large power would likely cause practically insignificant rejections of Ho.

This thesis shows the importance of understanding how sample size can affect the power of the two-sample Kolmogorov Smirnov test and can possibly lead to statistically significant results when no practical difference is actually present. It is important that a researcher does not conclude that there is a significant difference between two samples when the actual difference is only due to the natural variation in the data. It may be advisable, when possible, for a researcher to consider the magnitude of difference that is practically significant for the data they are attempting to analyze before actually analyzing the data. They can then use simulations, or some other method, to estimate the power of their test before making inferences.

It may be possible to use subsampling procedures in order to obtain more practically significant results for the two-sample KS test. This method would be easy to implement if a researcher had a power plot similar to the ones created in this thesis and an understanding of the practical difference in the magnitude they would like to reject for. Of course, this would only work properly if the variances of the two samples could be considered identical and differences in the mean were the primary concern. It is clear that the power of a two-sample Kolmogorov-Smirnov test can become very large in cases of large sample sizes. A prudent researcher should consider this when using the two-sample Kolmogorov Smirnov test for analysis.

REFERENCES

Boyerinas , B. M. (2016, December). *Determining the Statistical Power of the Kolmogorov-Smirnov and Anderson-Darling Goodness-of-Fit Tests via Monte Carlo Simulation*. Retrieved from https://www.cna.org/CNA_files/PDF/DOP-2016-U-014638-Final.pdf

Capon, J. (1965). On the Asymptotic Efficiency of the Kolmogorov-Smirnov Test. *American Statistical Association,* 843-853.

Daniel, W. W. (1990). Kolmogorov-Smirnov Two-Sample Test. In W. W. Daniel, *Applied Nonparametric Statistics* (p. 331). Boston: Brooks/Cole Cengage Learning.

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results.* New York: Cambridge University Press.

Ellis, S. M., & Steyn, H. S. (2013, January). Management Dynamics : Journal of the Southern African Institute for Management Scientists, Volume 12, Issue 4. *Journal of the Southern African Institute for Management Scientists*, 51-53.

Hojat, M., & Xu, G. (2004). A Visitor's Guide to Effect Sizes - Statistical Significance Versus Practical (Clinical) Importance of Research Findings. *Methodologist's Corner*, 241-249.

Hopkins, D. G., & Steele, D. D. (2011, November 11). *Devils Lake Basin Water Utilization Test Project: Soil Sustainability Investigations.* North Dakota State University, Fargo.

Klotz, J. (2012). Asymptotic Efficiency of the Two-Sample Kolmogorov-Smirnov Test. *Journal of the American Statistical Association*, 932-938.

Massey Jr., F. J. (2012). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 68-78.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*.

Ramachandramurty, P. C. (1966). On the Pitman Efficiency of One-Sided Kolmogorov and Smirnov Tests for Normal Alternatives. *The Annals of Mathematical Statistics,* 940-944.

SAS Institute Inc. (2010, April 30). *Tests Based on the Empirical Distribution Function.* Retrieved from SAS/STAT(R) 9.2 User's Guide, Second Edition: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_npar1way_a0000000202.htm

SAS Institute Inc. (2016). *SAS/GRAPH® 9.4: Reference, Fifth Edition.* Retrieved from documentation.sas.com: http://documentation.sas.com/api/collections/pgmmvacdc/9.4/docsets/graphref/content/graphref.pdf?locale=en#nameddest=n1e0onjgevom4kn1hbiibc6w34oe

Veris Technologies, Inc. (2014). *What drives soil EC, drives productivity*. Retrieved from veristech: https://www.veristech.com/the-soil/soil-ec

Yu, C. S. (1971). Pitman Efficiencies of Kolmogorov-Smirnov Tests. *The Annals of Mathematical Statistics*, 1595-1605.

APPENDIX A. TYPE ONE ERROR PLOTS

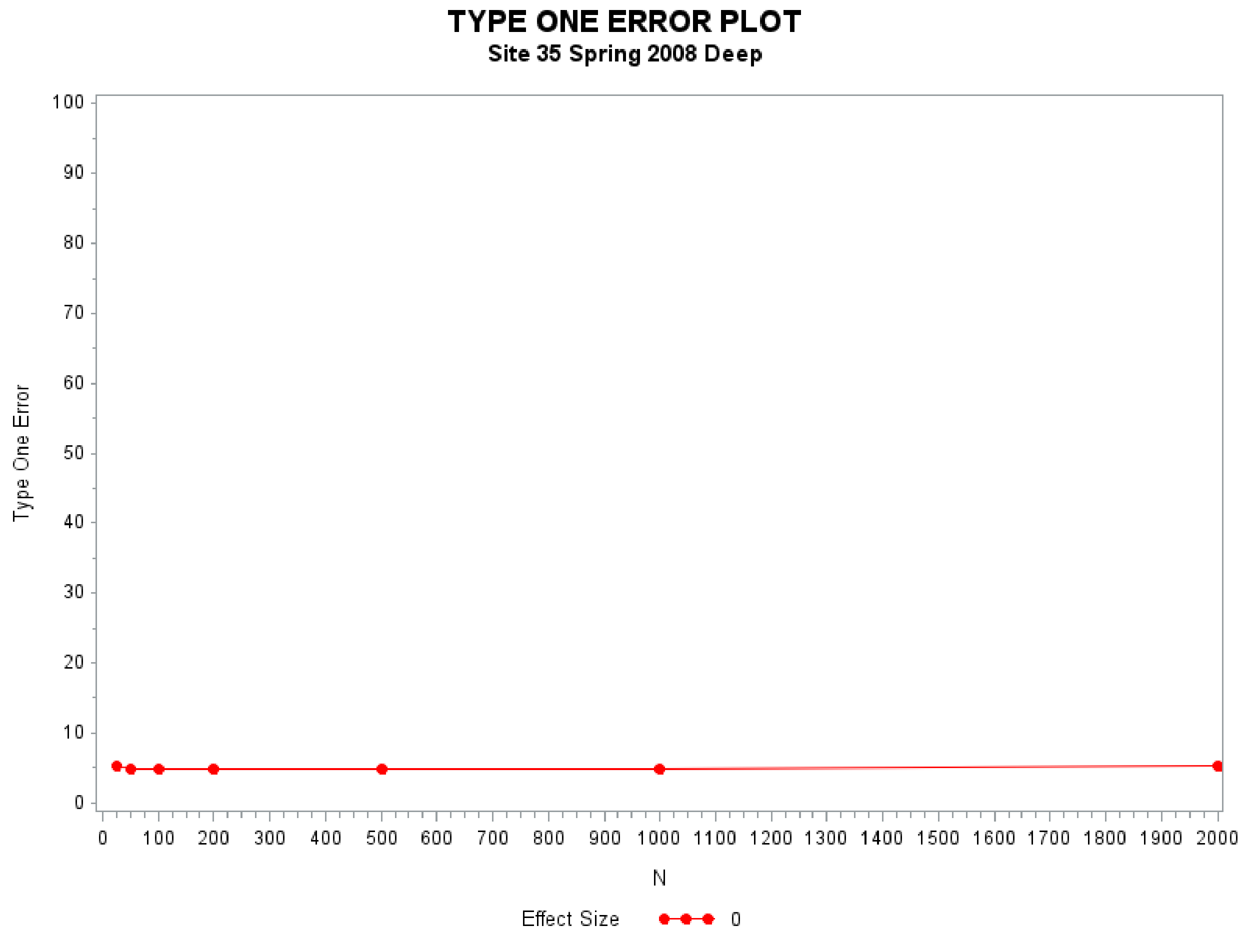

Figure A1. Site 35 Spring 2008 Deep Type One Error

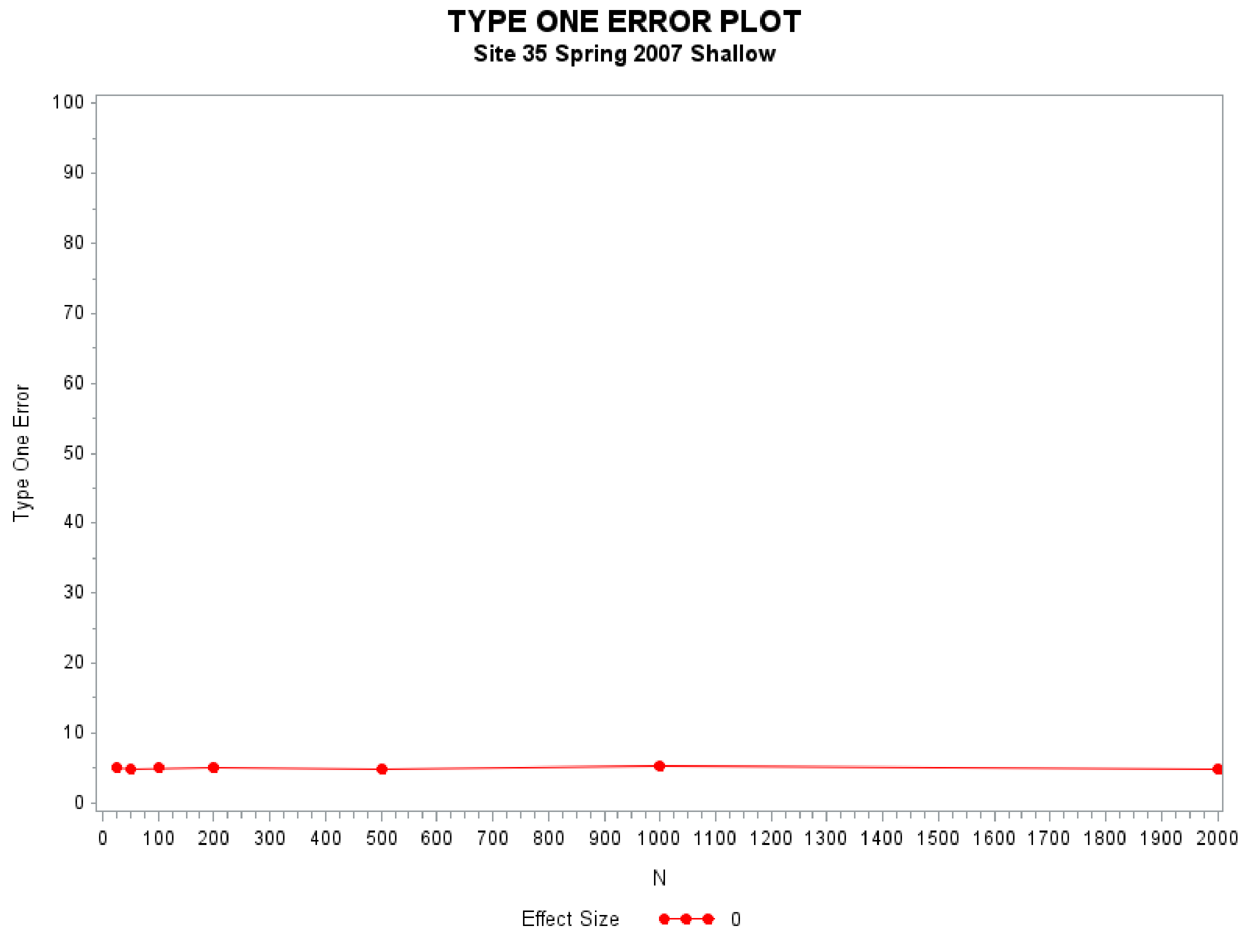

Figure A2. Site 35 Spring 2007 Shallow Type One Error

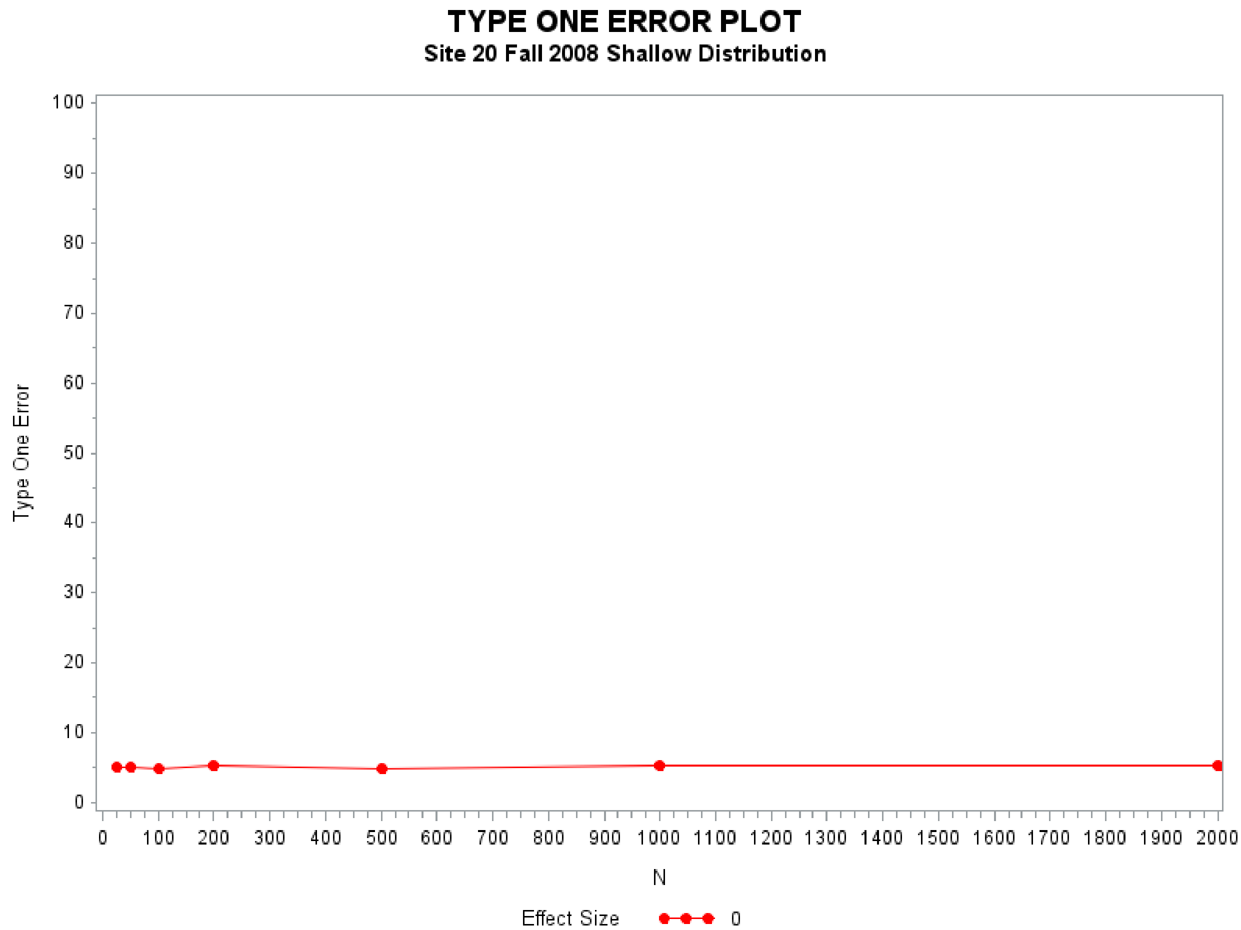

Figure A3. Site 20 Fall 2008 Shallow Type One Error

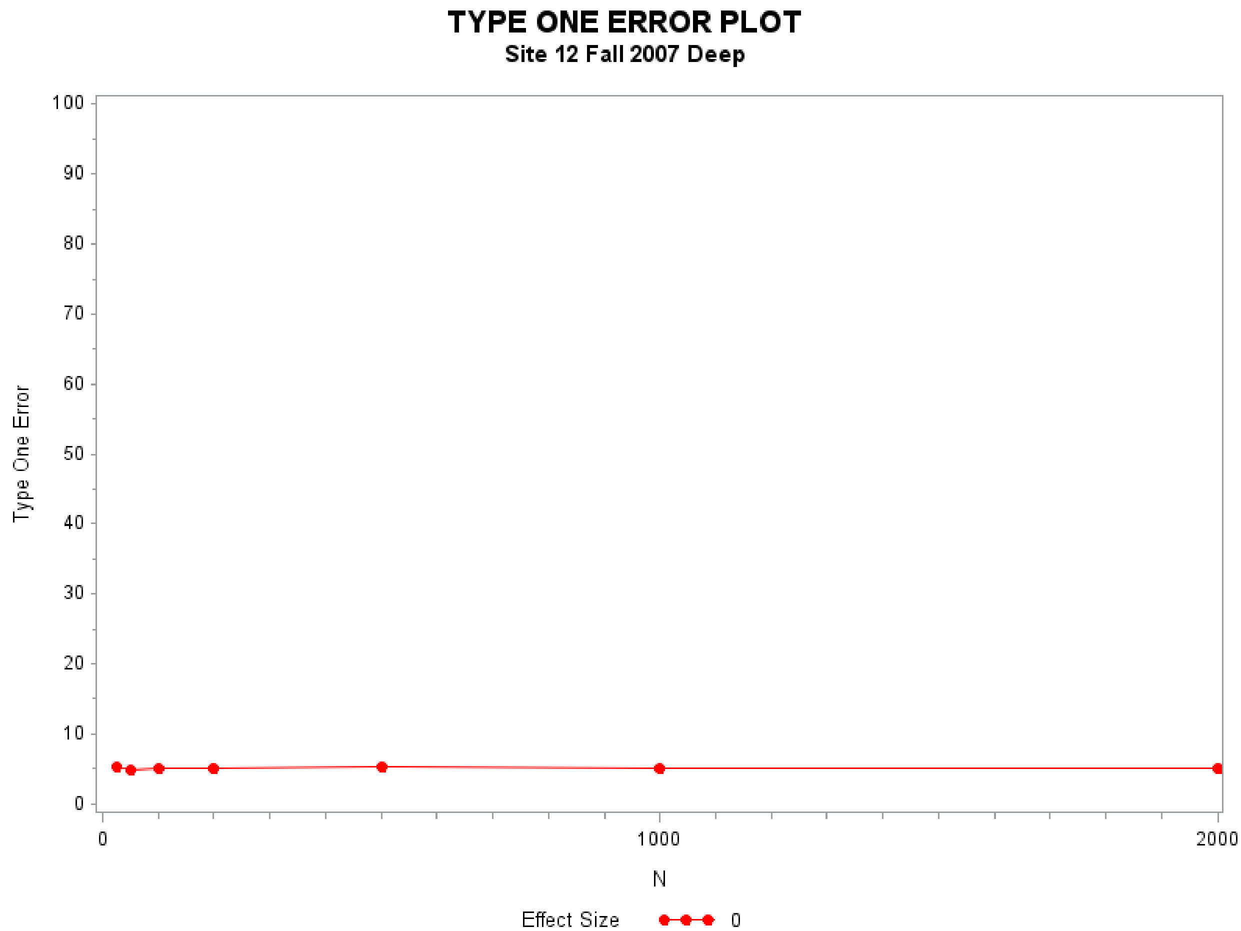

Figure A4. Site 12 Fall 2007 Deep Type One Error

APPENDIX B. POWER TABLES

Table B1. Site 35 Spring 2008 Deep Power Table

Site 35 Spring 2008 Deep		
Effect Size	Sample Size	Rejection Rate / Power
0	25	5.29%
0	50	4.89%
0	100	4.79%
0	200	4.82%
0	500	4.80%
0	1000	4.88%
0	2000	5.32%
0.1	25	13.24%
0.1	50	27.67%
0.1	100	68.63%
0.1	200	97.27%
0.1	500	100.00%
0.1	1000	100.00%
0.1	2000	100.00%
0.2	25	43.96%
0.2	50	82.12%
0.2	100	99.71%
0.2	200	100%
0.2	500	100%
0.2	1000	100%
0.2	2000	100%
0.5	25	92.89%
0.5	50	99.99%
0.5	100	100%
0.5	200	100%
0.5	500	100%
0.8	25	98.87%
0.8	50	100%
0.8	100	100%
0.8	200	100%

Table B2. Site 35 Spring 2007 Shallow Power Table

Site 35 Spring 2007 Shallow		
Effect Size	Sample Size	Rejection Rate / Power
0	25	4.99%
0	50	4.95%
0	100	4.99%
0	200	5.12%
0	500	4.94%
0	1000	5.28%
0	2000	4.80%
0.1	25	44.17%
0.1	50	81.26%
0.1	100	99.56%
0.1	200	100%
0.1	500	100%
0.1	1000	100%
0.1	2000	100%
0.2	25	91.86%
0.2	50	99.95%
0.2	100	100%
0.2	200	100%
0.2	500	100%
0.2	1000	100%
0.2	2000	100%
0.5	25	99.81%
0.5	50	100%
0.5	100	100%
0.5	200	100%
0.8	25	99.98%
0.8	50	100%
0.8	100	100%
0.8	200	100%

Table B3. Site 20 Fall 2008 Shallow Power Table

Site 20 Fall 2008 Shallow		
Effect Size	Sample Size	Rejection Rate / Power
0	25	5.40%
0	50	5.12%
0	100	4.86%
0	200	5.21%
0	500	4.91%
0	1000	5.29%
0	2000	5.31%
0.1	25	6.56%
0.1	50	7.86%
0.1	100	11.43%
0.1	200	18.30%
0.1	500	37.62%
0.1	1000	65.66%
0.1	2000	93.42%
0.2	25	10.13%
0.2	50	15.31%
0.2	100	28.09%
0.2	200	50.12%
0.2	500	89.05%
0.2	1000	99.62%
0.2	2000	100%
0.5	25	37.72%
0.5	50	65.29%
0.5	100	92.75%
0.5	200	99.90%
0.5	500	100%
0.8	25	75.65%
0.8	50	96.64%
0.8	100	99.98%
0.8	200	100%

Table B4. Site 12 Fall 2007 Deep Power Table

Site 12 Fall 2007 Deep		
Effect Size	Sample Size	Rejection Rate / Power
0	25	5.25%
0	50	4.91%
0	100	4.96%
0	200	5.17%
0	500	5.35%
0	1000	4.99%
0	2000	5.08%
0.1	25	6.69%
0.1	50	8.21%
0.1	100	10.64%
0.1	200	17.43%
0.1	500	37.99%
0.1	1000	67.95%
0.1	2000	95.90%
0.2	25	10.48%
0.2	50	16.53%
0.2	100	31.70%
0.2	200	56.81%
0.2	500	94.88%
0.2	1000	100%
0.2	2000	100%
0.5	25	43.81%
0.5	50	75.35%
0.5	100	98.57%
0.5	200	100%
0.5	500	100%
0.8	25	84.80%
0.8	50	99.19%
0.8	100	100%
0.8	200	100%

Table B5. Site 35 Spring 2007 Shallow with Actual Effects

Site 35 Spring 2007 Shallow for Actual Effects		
Actual Effect	Sample Size	Rejection Rate / Power
0.93	25	5.17%
0.93	50	5.68%
0.93	100	5.98%
0.93	200	7.41%
0.93	500	10.67%
0.93	1000	18.26%
0.93	2000	37.60%
1.87	25	5.41%
1.87	50	6.21%
1.87	100	9.35%
1.87	200	14.17%
1.87	500	30.79%
1.87	1000	64.52%
1.87	2000	94.67%
4.67	25	10.08%
4.67	50	16.45%
4.67	100	38.76%
4.67	200	71.67%
4.67	500	99.41%
4.67	1000	100%
7.47	25	20.96%
7.47	50	42.11%
7.47	100	82.23%
7.47	200	99.09%

APPENDIX C. ADDITIONAL INFORMATION ON SELECTED SITES

C.1. Site 35 Spring 2008 Deep

Site 35 2008 Deep Spring

The UNIVARIATE Procedure
Variable: Deep

Moments

N	3077	**Sum Weights**	3077
Mean	102.245141	**Sum Observations**	314608.3
Std Deviation	77.7849869	**Variance**	6050.50419
Skewness	0.81812368	**Kurtosis**	-0.6845994
Uncorrected SS	50778521	**Corrected SS**	18611350.9
Coeff Variation	76.0769518	**Std Error Mean**	1.40227127

Figure C.1. Site 35 Spring 2008 Deep Summary Statistics

C.2. Site 35 Spring 2007 Shallow

Site 35 2007 Shallow Spring

The UNIVARIATE Procedure
Variable: Shallow

Moments

N	2917	**Sum Weights**	2917
Mean	110.765924	**Sum Observations**	323104.2
Std Deviation	114.397093	**Variance**	13086.695
Skewness	1.70159438	**Kurtosis**	2.19174881
Uncorrected SS	73949737.7	**Corrected SS**	38160802.5
Coeff Variation	103.278237	**Std Error Mean**	2.11810154

Figure C.2. Site 35 Spring 2007 Shallow Summary Statistics

C.3. Site 20 Spring 2008 Shallow

Site 20 2008 shallow fall

The UNIVARIATE Procedure
Variable: Shallow

Moments

N	3254	**Sum Weights**	3254
Mean	28.8161647	**Sum Observations**	93767.8
Std Deviation	9.33122425	**Variance**	87.071746
Skewness	1.07415302	**Kurtosis**	4.82832509
Uncorrected SS	2985272.76	**Corrected SS**	283244.39
Coeff Variation	32.381909	**Std Error Mean**	0.16357987

Figure C.3. Site 20 Spring 2008 Shallow Summary Statistics

C.4. Site 12 Fall 2007 Deep

Site 12 Fall 2007 Deep

The UNIVARIATE Procedure
Variable: Deep

Moments

N	15906	**Sum Weights**	15906
Mean	52.6025525	**Sum Observations**	836696.2
Std Deviation	37.8623627	**Variance**	1433.55851
Skewness	1.09918963	**Kurtosis**	1.34742676
Uncorrected SS	66813103.9	**Corrected SS**	22800748.1
Coeff Variation	71.9781854	**Std Error Mean**	0.30021143

Figure C.4. Site 12 Fall 2007 Deep Summary Statistics

APPENDIX D. SAS CODE

D.1. Example Null Distribution Code

```
%let samples=100000;        /*How many samples (100000 for reference distribution) */
%let seed0=0;          /*Seed (set default to zero)*/
%let n=;              /*Sample size*/

sasfile Work.site20fall08 load;
data abc.gen_data (keep=sample iter trt shallow);

  call streaminit(&seed0);  *** Initialize with desired seed. ***;

  do sample=1 to &samples;
   do iter=1 to &n+1;
     trt='A';
     p = ceil(NObs * rand("Uniform"));    /* random integer 1-NObs */
     set Work.site20fall08 nobs=NObs point=p;  /* 2. POINT= observation; */
     output;
   end;
  do iter=1 to &n;
     trt='B';
     p = ceil(NObs * rand("Uniform"));    /* random integer 1-NObs */
     set Work.site20fall08 nobs=NObs point=p;  /* 2. POINT= observation; */
          shallow=shallow;
     output;
   end;
  end;
  STOP;
  run;
sasfile Work.site20fall08 close;

ods graphics off;  ods exclude all;  ods noresults;
*ods trace on;
ods output KSTest=KS_out
        KS2Stats=abc.site20reference&n. (where=(Name1='_KSA_'));
proc npar1way data=abc.gen_data  edf;
  by sample;
  class trt;
  var shallow;
  run;
```

D.2. Example Percentile Calculation Code

```
Proc Univariate data=site20reference&n.;
 var nValue1;
 Histogram nValue1;
 output out=qtls pctlpre=P_ pctlpts=1 to 99 by 1;
Run;

Proc transpose data=qtls out=UVqtls(rename=(Col1=KSD_value)) name=Percentile;
RUN;

proc print data=UVqtls;
Run;
```

D.3. Example Sample and Test Statistic Code

```
%let samples=10000;         /*How many samples (10000 for samples) */
%let seed0=0;           /*Seed (set default to zero)*/
%let n=;              /*Sample size*/
%let effectadd=0;      /*How much to add to trt B in order to get desired effect size*/

sasfile Work.site20fall08 load;
data abc.gen_data (keep=sample iter trt shallow);

 call streaminit(&seed0);  *** Initialize with desired seed. ***;

 do sample=1 to &samples;
  do iter=1 to &n+1;
    trt='A';
    p = ceil(NObs * rand("Uniform"));    /* random integer 1-NObs */
    set Work.site20fall08 nobs=NObs point=p;  /* 2. POINT= observation; */
    output;
  end;
 do iter=1 to &n;
    trt='B';
    p = ceil(NObs * rand("Uniform"));    /* random integer 1-NObs */
    set Work.site20fall08 nobs=NObs point=p;  /* 2. POINT= observation; */
         shallow=shallow + &effectadd;
    output;
  end;
 end;
 STOP;
 run;
sasfile Work.site20fall08 close;

ods graphics off;  ods exclude all;  ods noresults;
*ods trace on;
ods output KSTest=KS_out
       KS2Stats=abc.site20effectpoint0&n. (where=(Name1='_KSA_'));
proc npar1way data=abc.gen_data  edf;
 by sample;
 class trt;
 var shallow;
```

```
  run;
ods graphics on;  ods exclude none;  ods results;
```

D.4. Example Combining Code

```
/*effect=0.8*/
Title1 '0.8 effect size';
DATA site20shalloweffectpoint8;
 SET site20effectpoint825 (IN=N1)
     site20effectpoint850 (IN=N2)
     site20effectpoint8100 (IN=N3)
     site20effectpoint8200 (IN=N4);
 IF N1 THEN N=25;
  ELSE IF N2 THEN N=50;
  ELSE IF N3 THEN N=100;
  ELSE IF N4 THEN N=200;

 RUN;

/*effect=0.5*/
Title1 '0.5 effect size';
DATA site20shalloweffectpoint5;
 SET site20effectpoint525 (IN=N1)
     site20effectpoint550 (IN=N2)
     site20effectpoint5100 (IN=N3)
     site20effectpoint5200 (IN=N4)
     site20effectpoint5500 (IN=N5)
     site20effectpoint51000 (IN=N6);


 IF N1 THEN N=25;
  ELSE IF N2 THEN N=50;
  ELSE IF N3 THEN N=100;
  ELSE IF N4 THEN N=200;
  ELSE IF N5 THEN N=500;
  ELSE IF N6 THEN N=1000;

 RUN;


 /*effect=0.2*/
Title1 '0.2 effect size';
DATA site20shalloweffectpoint2;
 SET site20effectpoint225 (IN=N1)
     site20effectpoint250 (IN=N2)
     site20effectpoint2100 (IN=N3)
     site20effectpoint2200 (IN=N4)
     site20effectpoint2500 (IN=N5)
     site20effectpoint21000 (IN=N6)
     site20effectpoint22000 (IN=N7);
 IF N1 THEN N=25;
  ELSE IF N2 THEN N=50;
```

```
  ELSE IF N3 THEN N=100;
  ELSE IF N4 THEN N=200;
  ELSE IF N5 THEN N=500;
  ELSE IF N6 THEN N=1000;
  ELSE IF N7 THEN N=2000;
 RUN;




/*effect=0.1*/
Title1 '0.1 effect size';
DATA site20shalloweffectpoint1;
 SET site20effectpoint125 (IN=N1)
     site20effectpoint150 (IN=N2)
     site20effectpoint1100 (IN=N3)
     site20effectpoint1200 (IN=N4)
     site20effectpoint1500 (IN=N5)
     site20effectpoint11000 (IN=N6)
     site20effectpoint12000 (IN=N7);
 IF N1 THEN N=25;
  ELSE IF N2 THEN N=50;
  ELSE IF N3 THEN N=100;
  ELSE IF N4 THEN N=200;
  ELSE IF N5 THEN N=500;
  ELSE IF N6 THEN N=1000;
  ELSE IF N7 THEN N=2000;
 RUN;


 /*effect=0*/

Title1 '0 effect size';
DATA site20shalloweffectpoint0;
 SET site20effectpoint025 (IN=N1)
     site20effectpoint050 (IN=N2)
     site20effectpoint0100 (IN=N3)
     site20effectpoint0200 (IN=N4)
     site20effectpoint0500 (IN=N5)
     site20effectpoint01000 (IN=N6)
     site20effectpoint02000 (IN=N7);
 IF N1 THEN N=25;
  ELSE IF N2 THEN N=50;
  ELSE IF N3 THEN N=100;
  ELSE IF N4 THEN N=200;
  ELSE IF N5 THEN N=500;
  ELSE IF N6 THEN N=1000;
  ELSE IF N7 THEN N=2000;
 RUN;
```

D.5. Example Power Plot Code

```
DATA ALL;
 SET site20shalloweffectpoint1 (IN=E1)
    site20shalloweffectpoint2 (IN=E2)
          site20shalloweffectpoint5(IN=E3)
          site20shalloweffectpoint8(IN=E4);

 IF E1 THEN E=.1;
  ELSE IF E2 THEN E=.2;
        ELSE IF E3 THEN E=.5;
        ELSE IF E4 THEN E=.8;

 IF N=25 & nValue1 >= 1.30718 then reject=1;
 Else reject=0;
 If N=50 & nValue1 >= 1.32021 then reject=1;
 If N=100 & nValue1 >= 1.30754 then reject=1;
 If N=200 & nValue1 >= 1.31993 then reject=1;
 If N=500 & nValue1 >= 1.31474 then reject=1;
 If N=1000 & nValue1 >= 1.32349 then reject=1;
 If N=2000 & nValue1 >= 1.31876 then reject=1;

RUN;

PROC SORT;
 BY N;
RUN;

PROC FREQ DATA=ALL;
 BY N E;
 Tables reject/OUT=Percent2;
 RUN;
TITLE1 'POWER OF EFFECT SIZES';
TITLE2 'Site 20 Fall 2008 Shallow Distribution';


axis2 label=(angle=90 "Power")(order=0 to 100 by 10) minor=(n=1);
legend1 label = ('Effect Size');

PROC GPLOT Data=Percent2;
 WHERE reject=1;
 PLOT Percent*N=E/Haxis=0 to 2000 by 100 Vaxis=axis2 legend=legend1;
 SYMBOL1 V=DOT I=JOIN C=RED;
 SYMBOL2 V=PLUS I=JOIN C=BLACK;
 SYMBOL3 V=Triangle I=JOIN C=GREEN;
 SYMBOL4 V= CIRCLE I=JOIN C=BLUE;
 RUN;
```