Spreadsheets in Education (eJSiE)

Volume 4 | Issue 3 Article 4

6-21-2010

Bootstrapping Analysis, Inferential Statistics and EXCEL

John A. Rochowicz Jr
Alvernia University, john.rochowicz@alvernia.edu

Recommended Citation
Rochowicz, John A. Jr (2010) Bootstrapping Analysis, Inferential Statistics and EXCEL, *Spreadsheets in Education (eJSiE)*: Vol. 4: Iss. 3, Article 4.
Available at: http://epublications.bond.edu.au/ejsie/vol4/iss3/4

This Regular Article is brought to you by the Bond Business School at ePublications@bond. It has been accepted for inclusion in Spreadsheets in Education (eJSiE) by an authorized administrator of ePublications@bond. For more information, please contact Bond University's Repository Coordinator.

Bootstrapping Analysis, Inferential Statistics and EXCEL

Abstract
Performing a parametric statistical analysis requires the justification of a number of necessary assumptions. If assumptions are not justified research findings are inaccurate and in question. What happens when assumptions are not or cannot be addressed? When a certain statistic has no known sampling distribution what can a researcher do for statistical inference? Options are available for answering these questions and conducting valid research. This paper provides various numerical approximation techniques that can be used to analyze data and make inferences about populations from samples. The application of confidence intervals to inferential statistics is addressed. The analysis of data that is parametric as well as nonparametric is discussed. Bootstrapping analysis for inferential statistics is shown with the application of the Index Function and the use of macros and the Data Analysis Toolpak on the EXCEL spreadsheet. A variety of interesting observations are described.

Keywords
Bootstrapping, Resampling, Statistics, Inferences, Approximation

Bootstrapping Analysis, Inferential Statistics and EXCEL

John A. Rochowicz Jr

Alvernia University

john.rochowicz@alvernia.edu

Abstract

Performing a parametric statistical analysis requires the justification of a number of necessary assumptions. If assumptions are not justified research findings are inaccurate and in question. What happens when assumptions are not or cannot be addressed? When a certain statistic has no known sampling distribution what can a researcher do for statistical inference? Options are available for answering these questions and conducting valid research. This paper provides various numerical approximation techniques that can be used to analyze data and make inferences about populations from samples. The application of confidence intervals to inferential statistics is addressed. The analysis of data that is parametric as well as nonparametric is discussed. Bootstrapping analysis for inferential statistics is shown with the application of the Index Function and the use of macros and the Data Analysis Toolpak on the EXCEL spreadsheet. A variety of interesting observations are described.

Keywords: EXCEL, Spreadsheets, Inferential Statistics, Bootstrapping, Resampling Data, Numerical Approximations.

1. Introduction

EXCEL is prevalent, easy to learn and can be applied for numerous statistical projects. EXCEL, part of the Microsoft Office software package, has uses in a variety of educational and research settings including many for statistics and mathematics.

From my various classroom experiences, students doing quantitative research for undergraduate degrees in social work, psychology, and science to graduate degrees such as the MBA and PhD program in leadership and social sciences, apply most parametric tests of hypotheses without checking assumptions. Many examples of this lack of thoroughness on the part of the researcher have been observed from my personal experiences of working with faculty colleagues and mentoring research students. When parametric research is conducted assumptions must be met. Many students and faculty colleagues doing research do not address the assumptions required to conduct any of the typical parametric inferential statistical methods including t-tests, Analysis of Variances tests of hypothesis (ANOVA's) and regression analysis. As a result, testing hypotheses for inferential statistics are never completed and the reported results are open to debate.

2. Statistics: Parametric or Nonparametric

Parametric statistics are taught at all levels of post secondary education. Undergraduate to graduate levels and for all types of majors from the social to the physical sciences, statistics are studied and applied in all kinds of research. The assumptions associated with parametric tests of hypothesis [4] include: 1) Data are collected randomly and independently (tests of hypothesis for means, Analysis of Variances (ANOVA's) for means and regression analysis). 2) Data are collected from normally distributed populations (tests of hypothesis for means, ANOVA'S for means and regression analysis). 3) Variances are equal when comparing two or more groups (tests of hypothesis for means, ANOVA'S for means and regression analysis).

Many ways exist for checking assumptions. In order to check that sampled data is from a normally distributed population, histograms or stem-leaf plots are usually displayed. If they appear mound-shaped the assumption of normality is accepted as verified. Another way to verify normality is applying a nonparametric test of hypothesis, the Kolgomorov-Smirnov goodness of fit test of hypothesis for fitting data to a specific distribution. In this test of hypothesis the researcher needs to accept the null hypothesis that the data are normal. In this way the assumption of normality is justified. Levene's Test of Hypothesis of the Equality of Variances determines whether variances for populations being studied are equal. The need to accept the null hypothesis of equal variances

verifies that the assumption that variances are equal. Usually these tests are used in conjunction with conducting parametric statistical inferences. The manner in which the research experiment is defined and how data are collected would justify the requirements that data are collected randomly and independently.
Examples of where parametric statistics are used include:

a) The researcher is interested in studying if there is a difference in the means for scores achieved by students taking a mathematics course compared to students taking a social science course. This is an example of where a parametric independent samples t-test would be conducted
b) A researcher is looking for a linear relationship between the time a student puts into study and the final grade for a course. In this situation linear regression and correlation analysis would be performed.

If the data collected are not normal or the homogeneity or equality of variance assumption is violated, there are alternative ways to make inferences. When assumptions are not validated, published results become invalid and questionable. A researcher that does not justify assumptions or cannot determine a specific sampling distribution for a statistic must apply nonparametric distribution-free statistical techniques.
Nonparametric techniques exist where no assumptions are needed or no sampling distributions are found. For example the Kruskal-Wallis test of hypothesis is a nonparametric alternative to the one way ANOVA This nonparametric technique tests a hypothesis about the nature of three or more populations and whether they have identical distributions and if they differ with respect to location [7]. There are many others [4], [5], and [7].
Another concern is the inability of determining the sampling distribution for certain statistics. For example there is no sampling distribution that can be used for population medians [5]. Computing technology and bootstrapping can be used for inferential statistics where the sampling distribution of the statistic is unknown. Bootstrapping is a nonparametric, numerical application that can be applied in EXCEL.

3. Bootstrapping Analysis: Nonparametric Analysis

The analysis of data without checking assumptions can be done with bootstrapping techniques. Bradley Efron [2] described a computer intensive technique that resamples collected data in order to study the behavior of a distribution of a specific statistic. As a result inferences are made on populations that are parametric as well as nonparametric.

Bootstrapping is a numerical sampling technique where the data sampled are resampled with replacement [2]. This means that you acquire a sample. Place sample values back and then select another sample. In this way you get sample data from which you can generate summary statistics for each resample. Various descriptive statistics such as mean, median, mode, variance and correlation can be bootstrapped.

With the use of a computer the student or researcher can create many resamples. Bootstrapping statistics allows the student or researcher to analyze any distribution and make inferences. The sampled data becomes the population and the resampled data are the samples. There are advantages as well as disadvantages when bootstrapping.

3.1 Advantages and Disadvantages:

Advantages include:

1) Verifying assumptions of normality and equality of variances for the population is unnecessary. Inferences are valid even when assumptions are not verified.

2) There is no need to determine the underlying sampling distribution for any population quantity.

3) Interpretations and results are based upon many observations.

Disadvantages include:

1) Powerful computers are necessary

2) Randomness must be understood

3) Computers have built-in error.

4) Large sample sizes must be generated.

4. Inferential Statistics: Confidence Intervals

In any parametric or nonparametric statistics the researcher's goal, is to infer something about the population. In order to make inferences about the population, constructing confidence intervals is acceptable. A confidence interval includes a sample statistic or point estimate plus or minus a constant error term. These ranges of values are found by setting the significance level (usually 0.05) for making a decision for the hypothesis about the population; determining the correct sampling distribution; and finding whether the theorized population value is in the confidence interval or not. The confidence interval takes on the form point estimate ± a critical value ($z_{\alpha/2}$) times the standard error (SE) of the statistic. For the mean, the confidence interval looks like

xbar ± $z_{\alpha/2}$ SE(sample mean), where xbar is the sample mean.

Testing hypothesis using confidence intervals is accomplished by checking whether the theorized population quantity is or is not in a certain confidence interval [1] and [4]. If the theorized population value is in the interval the null hypothesis is accepted and if not the null hypothesis is rejected. This means that the value obtained in the sample is not significantly different from the theorized population value if in the interval and the sample and population are significantly different when the population value is not in the interval. Suppose a researcher wants to find a 95% confidence interval for the mean. The error term is comprised of the standard error of the mean, that is the standard deviation over the square root of the sample size and a critical z for large samples (where population standard deviation is known) and t for small (where population standard deviation is unknown) samples. These z or t values are critical values and are based of the significance level set by a researcher. Alpha is the type I error or the probability of rejecting a true null hypothesis [4]. For a 95% confidence interval alpha is 1-.95 or 0.05. The values for $z_{\alpha/2}$ or $t_{\alpha/2}$ are found from z or t tables or in EXCEL with the application of the function "=TINV(probability, degrees of freedom)".

Consider finding a 95% confidence interval for the mean. Calculating such an interval indicates that upon repeated sampling 95% of the samples will contain the population mean. Suppose a researcher wishes to test the hypothesis that the population mean is 80 from a set of grades for 20 students in a certain statistics course. The data were 81 70 79 86 89 65 76 69 71 88 78 79 81 79 73 84 73 81 89 88. The sample mean was 79.02 and the standard deviation was 7.26. In order to test the hypothesis that sample mean is not significantly different from the theorized mean of 80, a 95% confidence interval is constructed. The researcher never knows what the population mean is, only an approximation. The researcher is 95% confident it's between the numbers that define the interval upon repeated sampling. That is the interval is in agreement with the sample the researcher obtained. If the population mean is outside that interval, then the sample mean is significantly different from the population mean.

The results of the calculations of the confidence interval for the mean in this example are as follows: The sample mean is 79.02. The confidence interval is comprised of the sample mean plus or minus $t_{\alpha/2}$ times the standard error of the mean. In this case, the t value based on an alpha of 0.05 and 20-1 degrees of freedom is 2.09 from a statistical table [4] and the standard error of the mean or the standard deviation divided by the square root of the sample size is 1.62. Combining these numbers provides the 95% confidence interval for the mean as $79.02 \pm (2.09)(1.62)$. The confidence interval is $75.62 < \mu < 82.42$, where μ is the population mean.

At the 5% or 0.05 significance level, the theorized mean of 80 is contained in the interval found and so the null hypothesis is not rejected. There is no significant difference between the sample mean of 79.02 and the population mean of 80. Example 1 shows how to use EXCEL to test this hypothesis about the mean.

If there is no known sampling distribution for a particular theorized population quantity such as median or log means [6] an alternative nonparametric confidence interval is found by using the resampled data and applying the 2.5 percentile and 97.5 percentile for the generated distribution In this way a 95% confidence interval is obtained from the lower 2.5 percentile and upper 97.5 percentile. These bootstrapped confidence intervals are used as any other confidence intervals to make inferences about the population [6].

The examples that follow show the application of EXCEL to inferential statistics and bootstrapping analysis. Classical confidence intervals and bootstrapped percentile confidence intervals are presented. Similar conclusions and results are reached. The determination of bootstrapped percentile confidence intervals is necessary for distributions of medians since there is no known sampling distribution .

5. EXCEL Applications

EXCEL is useful for doing parametric as well as nonparametric statistics. The simulation of normal data, the determination of t-values, means and medians for sets of data are described in the following examples. Also percentiles for sets of data can be found in EXCEL.

5.1 EXCEL: Classical Example

Consider the dataset of 20 statistics grades analyzed above. The classical way to make an inference concerning the mean is to: a) Identify the null and alternative hypothesis. b) Construct a confidence interval and c) Decide to reject or fail to reject the null hypothesis. The assumption that data are normal can be checked by checking a histogram. Using the fact that a histogram appears normal and data were generated using the function "=(NORMSINV(RAND())*standard deviation)+mean", the t-test of hypothesis for means can be applied. The rejection or failure to reject a null hypothesis is accomplished by using confidence intervals [4].

Figure 1 displays the classical method, the traditional textbook method for determining confidence intervals.

A class of 20 students took a statistics test and the class mean was 79.02 with a standard deviation of 7.26. A researcher wants to test whether the sample mean is significantly different from a theorized population mean of 80, an average grade necessary for meeting a statewide assessment mandate. In figure 1 the

results for a 95% confidence interval for the mean are shown. The confidence interval does contain the population mean of 80 and so the hypothesis of no difference between the population mean and the sample men is not rejected. Findings indicate there is no significant difference. And the $t_{\alpha/2}$ is found using "=tinv(probability, degrees of freedom)". Notice the value 2.09 agrees with the value from a textbook [4].

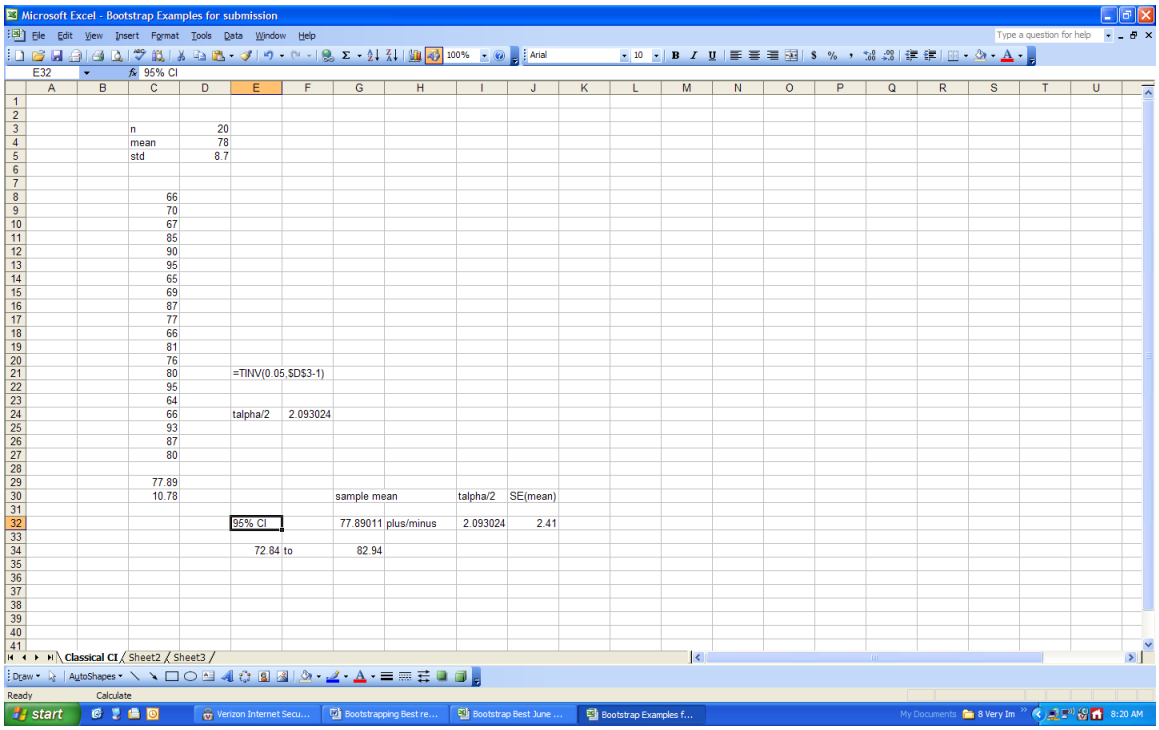

Figure 1: Classical t-test of Hypothesis with Confidence Intervals

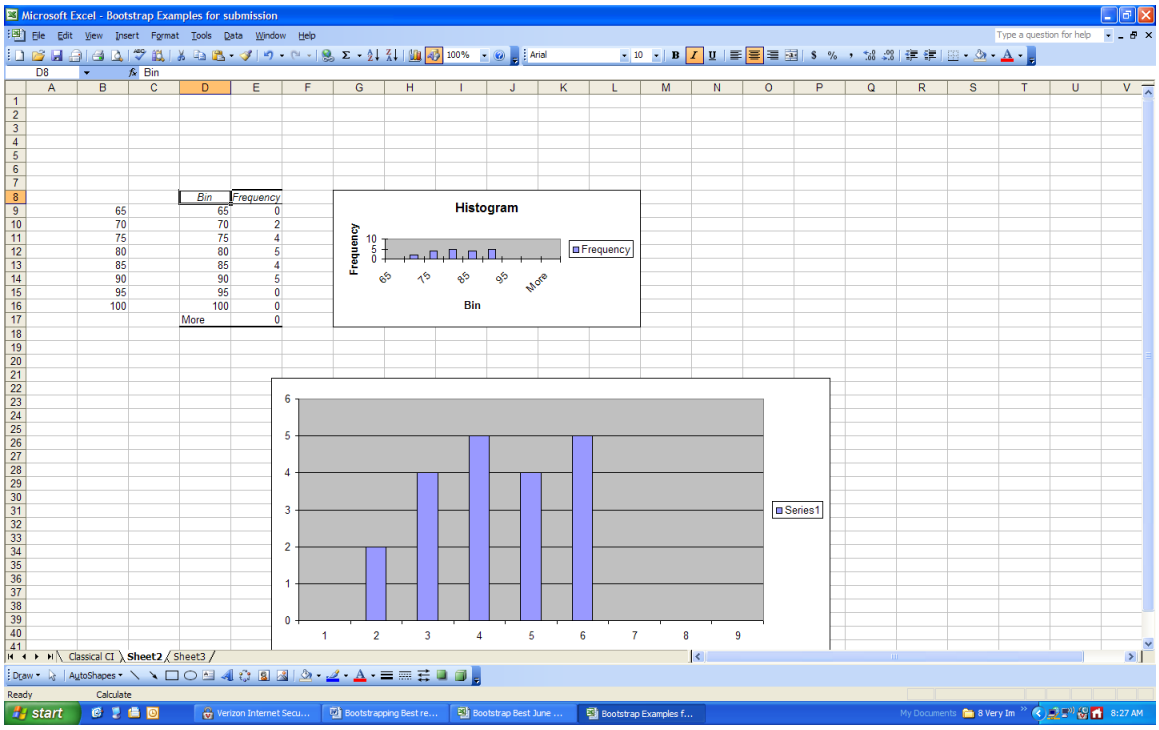

Figure 2: The Data are Normal.

The histogram appears mound shaped and the application of the t-distribution is appropriate. For any dataset the distribution would be normal since the function "=(NORMSINV(RAND())*standard deviation)+mean" was applied.

5.2 EXCEL EXAMPLE 2: Large Sample Statistics

The data for this example are based on IQ scores with a mean of 100 and a standard deviation of 15. A distribution of these IQ scores of 40 samples of size 30 is shown in figure 3. The distribution of means of the 40 samples with a histogram and frequency distribution are displayed in figure 4. The distribution of means appears to be mound shaped. For parametric statistics the shape of the sampling distribution is supposed to be approximately mound-shaped or normal. The data were simulated using the function "=NORMINV(RAND())*$C$2)+$C$1" where cell C2 contains the standard deviation and cell C1 contains the mean. This function is copy and pasted into as many cells as desired. In figure 5, the 95% confidence intervals for each sample and the distribution of means are illustrated. As expected about 95% of the classical confidence intervals for the mean do contain 100.

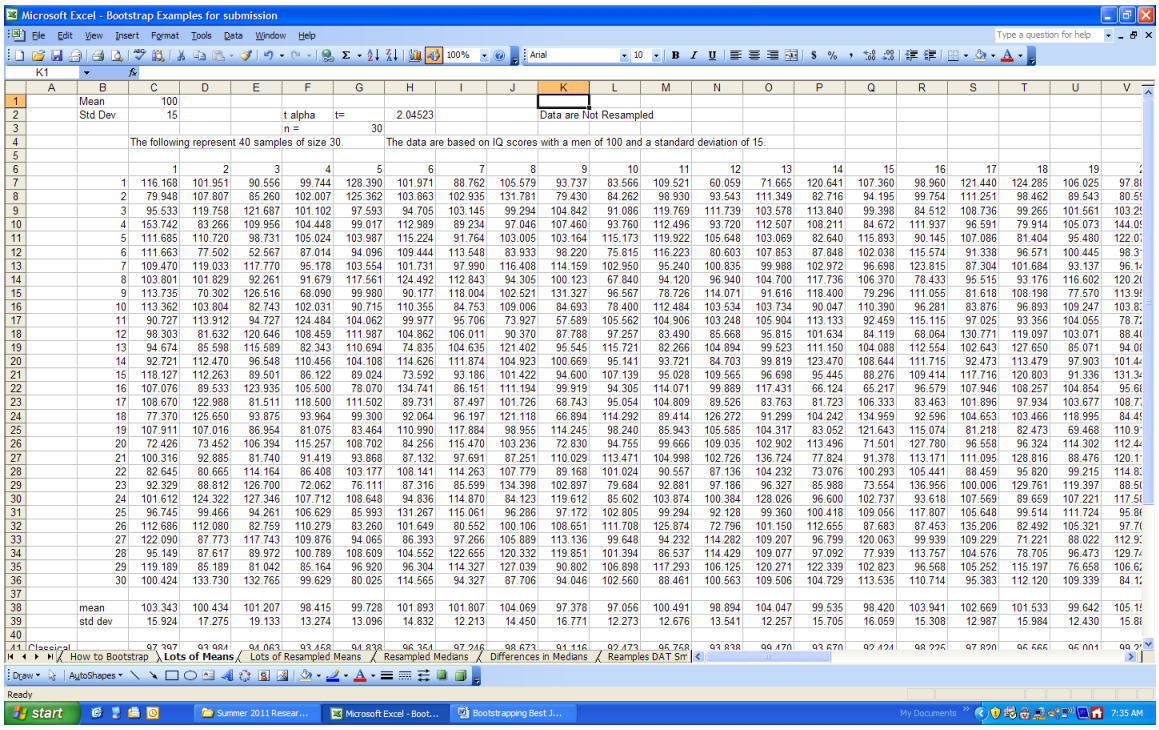

Figure 3: Section of 40 Samples

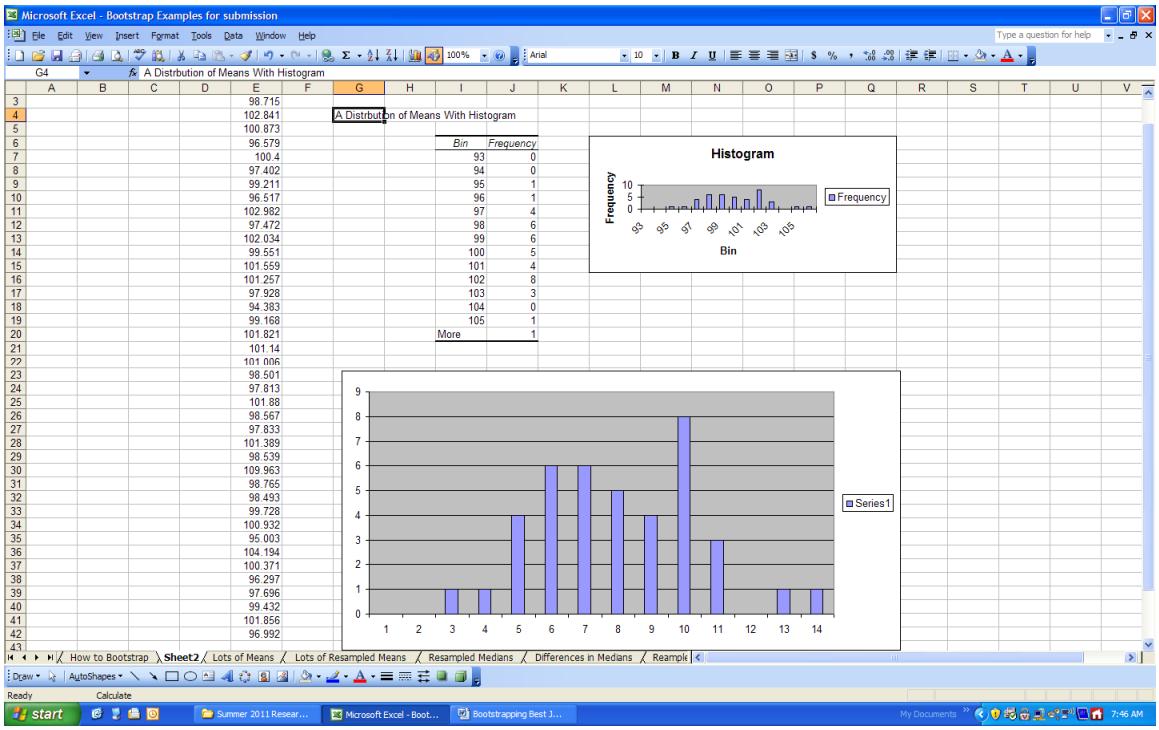

Figure 4: The Distribution of Means Appears Mound-shaped.

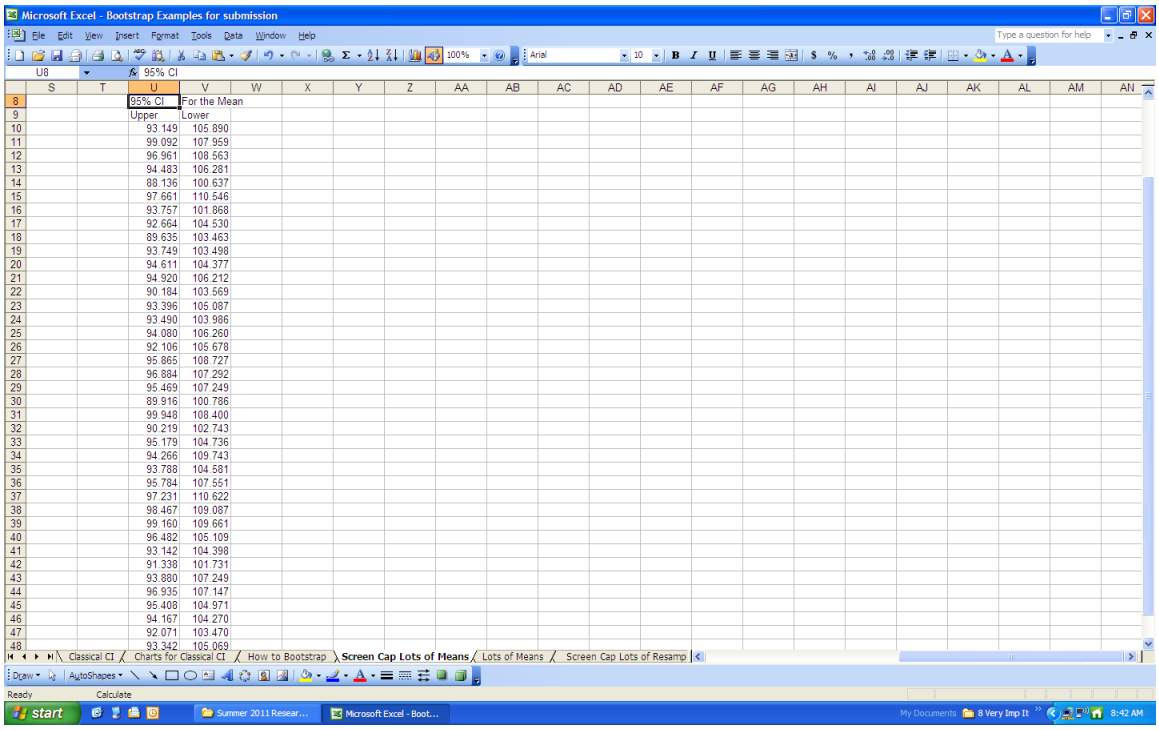

Figure 5: Classical Confidence Intervals for the Sample Means

Tests of hypothesis using confidence intervals can be conducted using EXCEL. If a theorized value is in a certain confidence interval accept the null hypothesis of no difference and if the theorized value is outside the interval reject the null hypothesis. The findings suggest that the population quantity is significantly different from the sample quantity acquired for a specific experiment. And the difference has not occurred by chance.

A hypothesis test can de performed on means by applying a 95% confidence interval for the mean using the classical t-distribution and the percentile method. In the case where no sampling distribution of a statistics is known constructing a 95% confidence interval using percentiles is applied. See examples 7 and 8 for percentile confidence intervals.

5.3 EXCEL Example 3: Bootstrapping a small sample

EXCEL does not have any built-in commands or programs to perform bootstrapping**.** But there are ways to do bootstrapping in EXCEL without the purchase and learning of other software such as SPSS. They include:

1) Applying the INDEX Function
2) The application of Data Analysis Toolpak and Macros. A detailed description of using the Data Analysis Toolpak and Macros for bootstrapping data is supplied in the Appendix

Applying the EXCEL INDEX Function is a way to conduct bootstrapping analysis without using an add-in or any macros. If you wish to do a large number of resamples, use the INDEX command and the random number generator. The primary use of the Index Function is to return a value from a table or range of data. The structure of the Index Function is: INDEX(table or a range or an array, row location, column location). The use of rand()+1 will generate a random row or column location. The syntax for the command that generates random rows and columns of data from a sample is the following: "=INDEX(range of cells, ROWS(range of cells)*RAND()+1,COLUMNS(range of cells)*RAND()+1)". Next copy and paste this command for as many resamples as desired. Pressing F9 key on computer keyboard recalculates data.

With the INDEX function, bootstrapping can be performed on any statistic including means, medians, modes, analyses of variance and regression analysis such as correlation and beta weights.

Consider the following set of data: 1 5 8 9 12 15 18. Using the function "=INDEX((c4:c10),ROWS(c4:c10*RAND()+1,COLUMNS(c4:c10*RAND()+1)" will generate random resamples. As expected some of the values are shown and others are not. Also some of the numbers occur more than once. The results are 18 9 5 8 1 8 8. Pressing F9 will show many resamples. Figure 6 shows one resample using "INDEX"

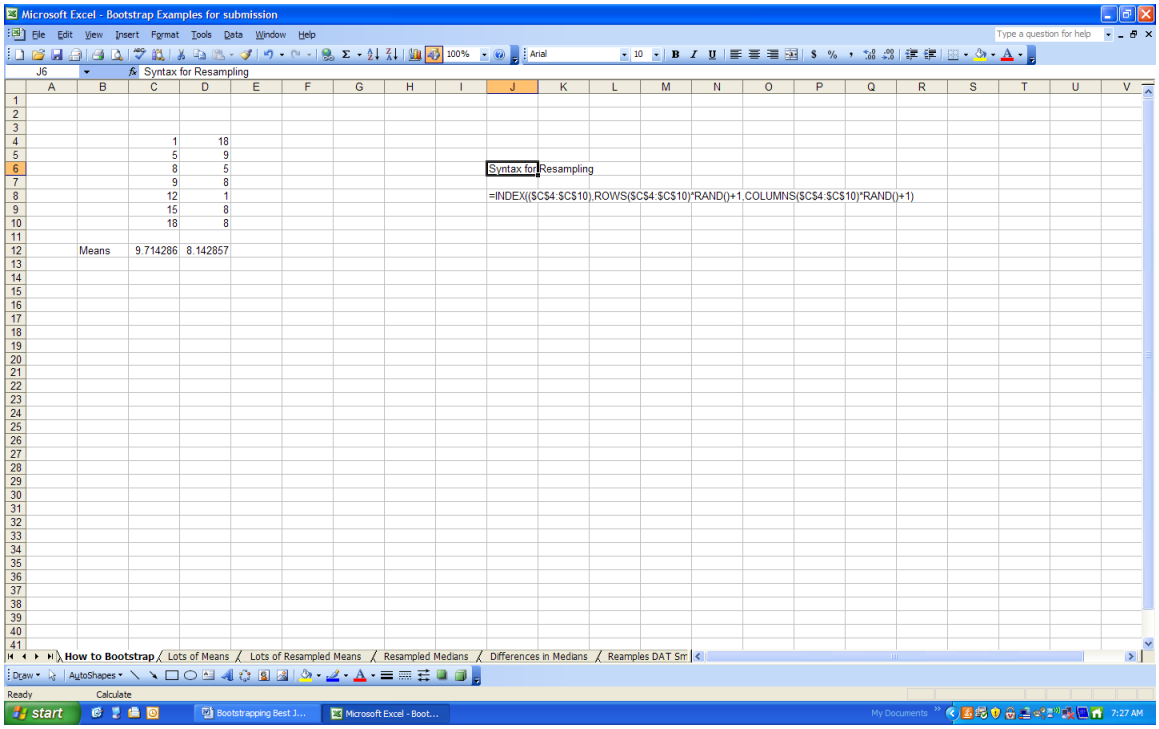

Figure 6: Resampling One Sample

5.4 EXCEL Example 4: Bootstrapping Analysis

Resample the data for 30 IQ scores from example 2 using the INDEX function. Forty resamples were obtained and displayed in figure 7. Notice that the distribution of means on the resampled data appears normal (figure 8). Since the data appears normal, the t-test of hypothesis can be applied and the classical confidence intervals can be determined.

Suppose you with to test whether the mean is not equal to 100, the mean for IQ scores for the population of all persons that take the IQ test.

With the use of confidence intervals a student or researcher can test this hypothesis.

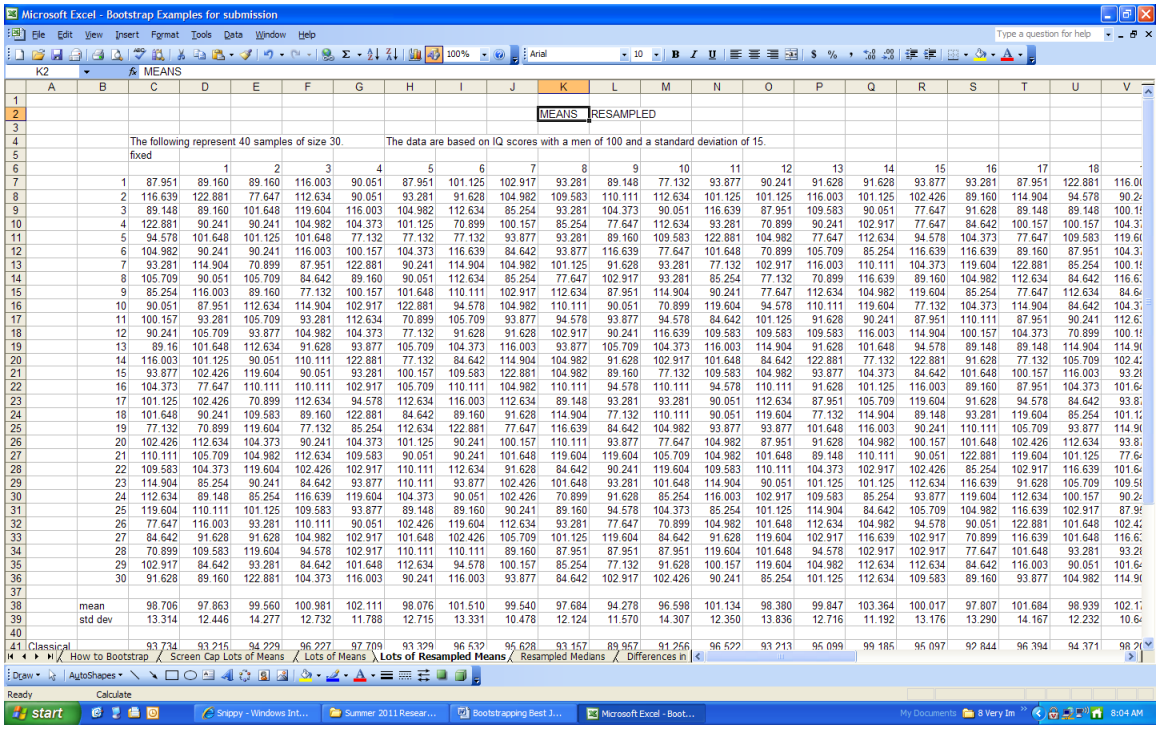

Figure 7: Section of Resampled Means

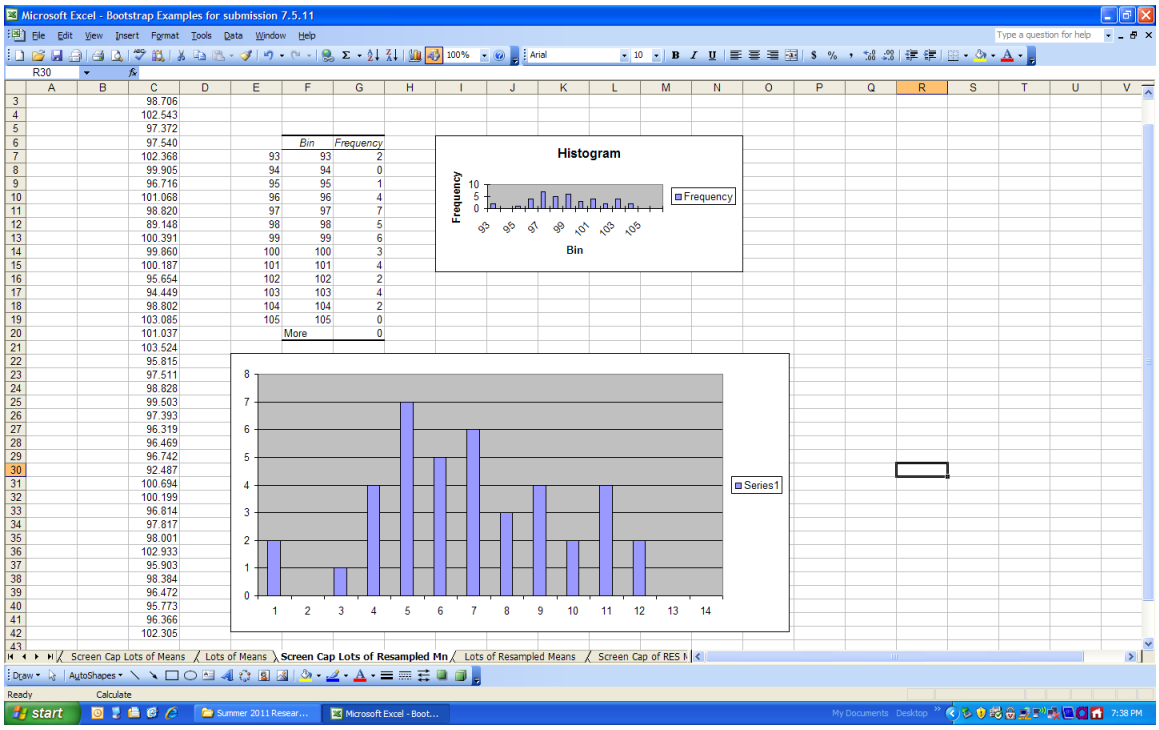

Figure 8: The Distribution of Resampled Means Appears Mound-Shaped

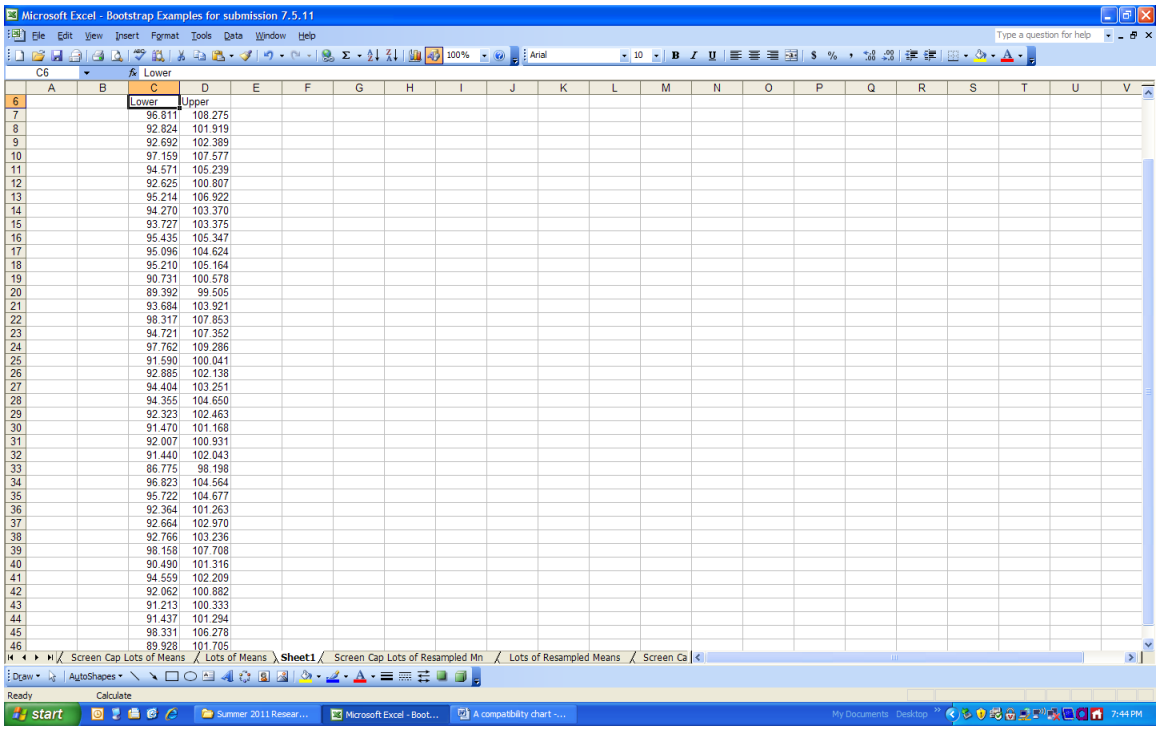

Figure 9: Confidence Intervals for Resampled Means

In figure 9 the classical confidence intervals for resampled means are presented. Using the distribution of means and the percentile functions "=PERCENTILE (range of cells, .025)" and "=PERCENTILE (range of cells, .975)" a 95% percentile confidence interval for the mean is found.

For figure 8 apply "=PERCENTILE (C3:C42, .025)" and "=PERCENTILE (C3:C42, .975)" to obtain the percentile confidence intervals. Using the distribution of resampled means of figure 8, the percentile confidence interval is 92.403 to 103.096. In the distribution of means with percentile confidence intervals, notice that results are very similar. That is about 95% of the confidence intervals for sample means contain the hypothesized mean of 100. And upon repeated resampling about 95% of the resamples contain 100.

When testing whether the population mean is 100, note that since 100 is in the interval the null hypothesis of no difference between the sample mean and the theorized mean of 100 is not rejected.

From figure 4 the confidence interval for means based on percentiles is 94.99 to 104.34.

5.5 EXCEL Example 5: Bootstrapping with the Data Analysis Toolpak

Check the appendix for installing Data Analysis Toolpak, applying the Data Analysis Toolpak and applying a macro in Data Analysis Toolpak. The following set of data 1 5 8 9 12 15 18 was resampled using the Data Analysis Toolpak. The results after one application are 18 18 12 15 9 15 8. Some data are the same and some are missing. Bootstrapping has occurred. Figure 10 shows the results of using the Toolpak.

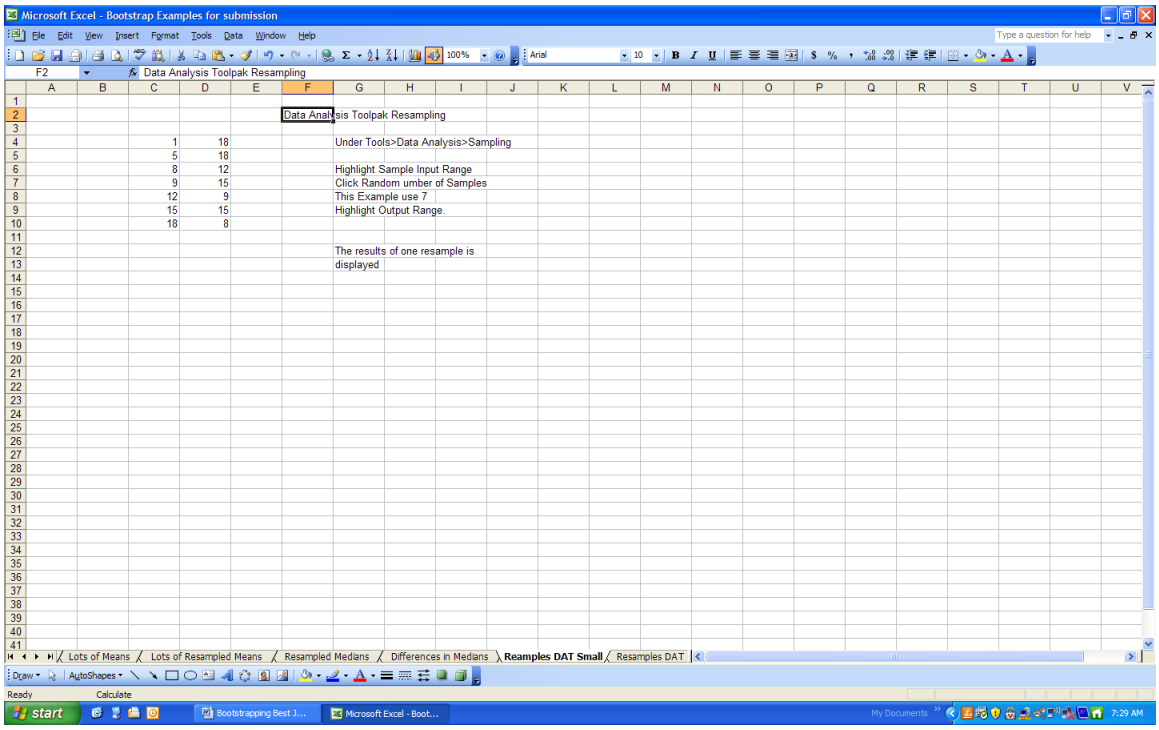

Figure 10: Resampling with Data Analysis Toolpak

5.6 EXCEL Example 6: Bootstrapping Analysis with Data Analysis Toolpak: Many Resamples

Since bootstrapping analysis requires many "resampled" samples the process can be repeated as many times as desired by changing the output range in the Sampling Dialog Box but this tends to be tedious In order to automate this process the development of a macro will help. The macro is in the appendix. The macro automates the calculation of many reseamples. The example above for bootstrapped means is analyzed by using the Toolpak. Figure 11 displays the results. The distribution appears mound-shaped and t-distribution confidence intervals can be constructed as in example 4.

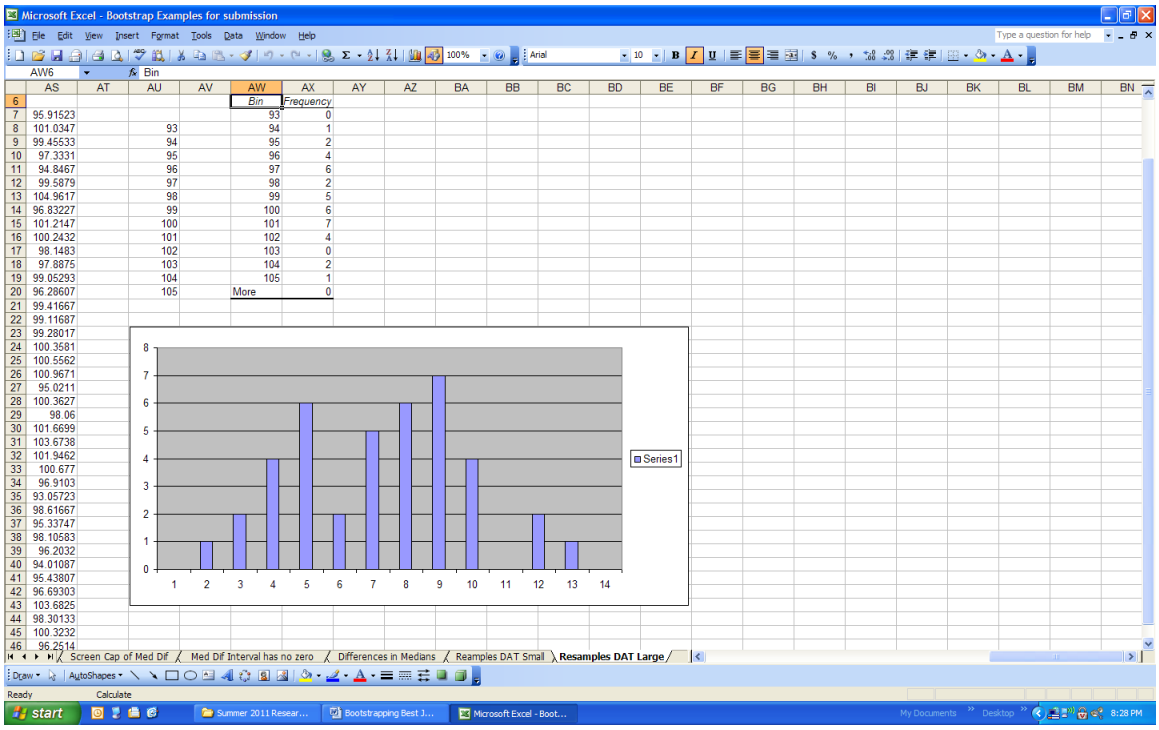

Figure 11: Distribution of Resampled Means with Data Analysis Toolpak

5.7 EXCEL Example 7: Inferential Statistics on Medians Using Bootstrapping Techniques

Consider a dataset of size 30. The data are 20 25 33 42 48 51 60 72 75 74 81 87 102 105 110 123 142 151 159 200 214 234 244 300 500 602 603 604 609 651.

Resample this data set 40 times and obtain a distribution of medians. Since there is no known sampling distribution for medians the application of bootstrapping techniques should be used. Generating a distribution of medians is accomplished by resampling with the INDEX function. Figure 12 displays a sample of size 30 resampled 40 times. The distribution of medians is not usually normal and since there is no known sampling distribution for medians bootstrapping analysis is applied. Calculating a 95% confidence interval for the distribution of medians is shown in figure 13. This confidence interval for the population median is based on percentiles and leads to decisions about significant differences in the 2 groups.

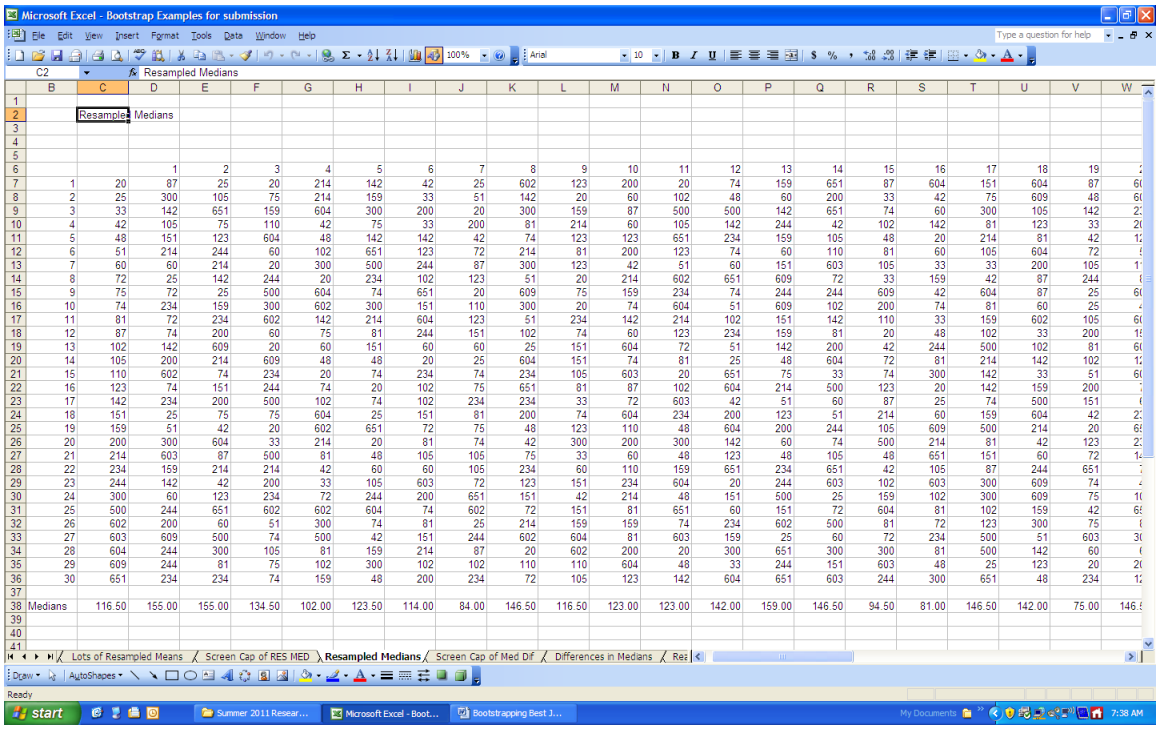

Figure 12: Section of Resampled Medians

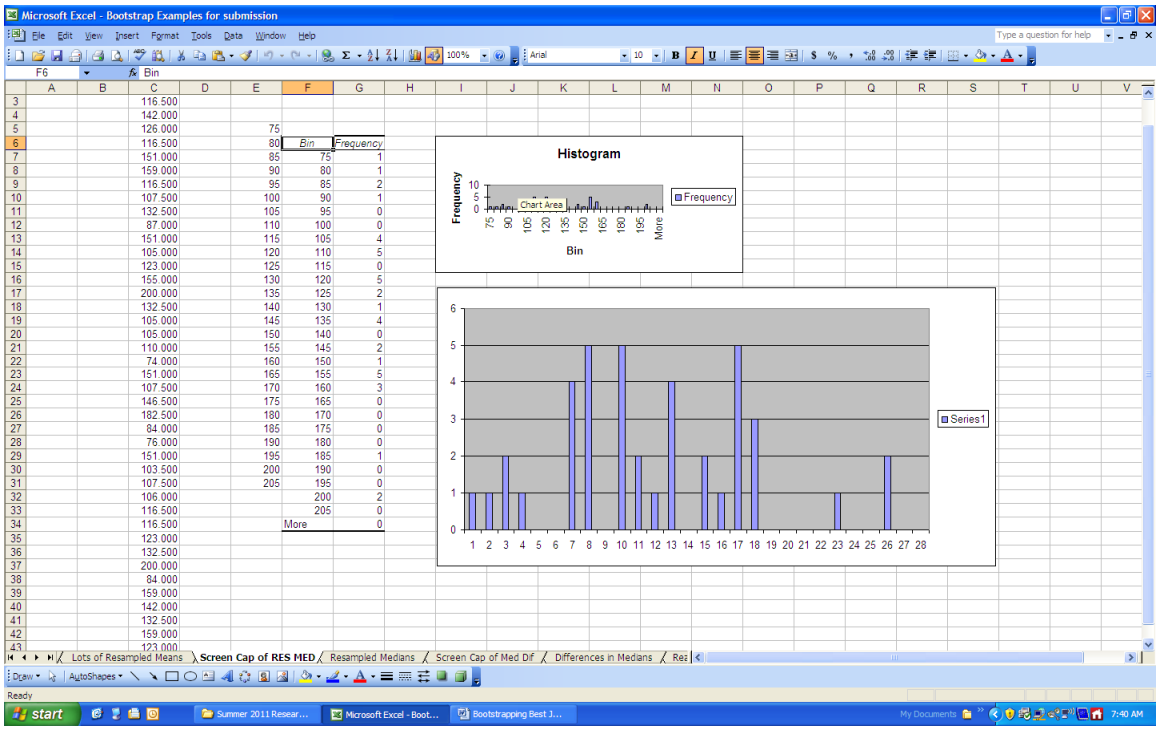

Figure 13: Distribution of Resampled Medians.

The histogram is not normal and so percentile confidence intervals are used to make an inference about the population median. The percentile confidence interval for the distribution of medians in figure 13 is 76 to 200. Suppose a researcher wants to test whether the null hypothesis for the population median is 150. Based on the percentile confidence interval the null hypothesis would not be rejected, since 150 is in the interval 76 to 200. Therefore there is no significant difference between 150 and the sample median of 126, the average of all resampled medians.

5.8 EXCEL Example 8: Inferential Statistics on the Differences in Medians

Suppose the following data represent the prices on homes in 2 different locations.
The prices on homes at location A are: 99 96 93 92 87 81 82 89 77 74 76 66 71 82 69 71 80 66 42 45 33 46 32 22 25 19 15 17 14 12
The prices on homes at location B are: 300 333 321 345 333 245 324 222 321 119 117 115 111 100 96 65 62 61 69 71 45 42 55 69 78 88 67 42 66 68
Consider the difference in the median prices of the homes in the 2 locations. Is there a significant difference in the median prices?
The first column is the sampled data for each dataset. The data is resampled 40 times for each set of data. Inferences are made on whether there is a significant difference in the median prices at the 2 locations based on percentile confidence intervals. That is the difference in the median prices at the 2 locations is 0.

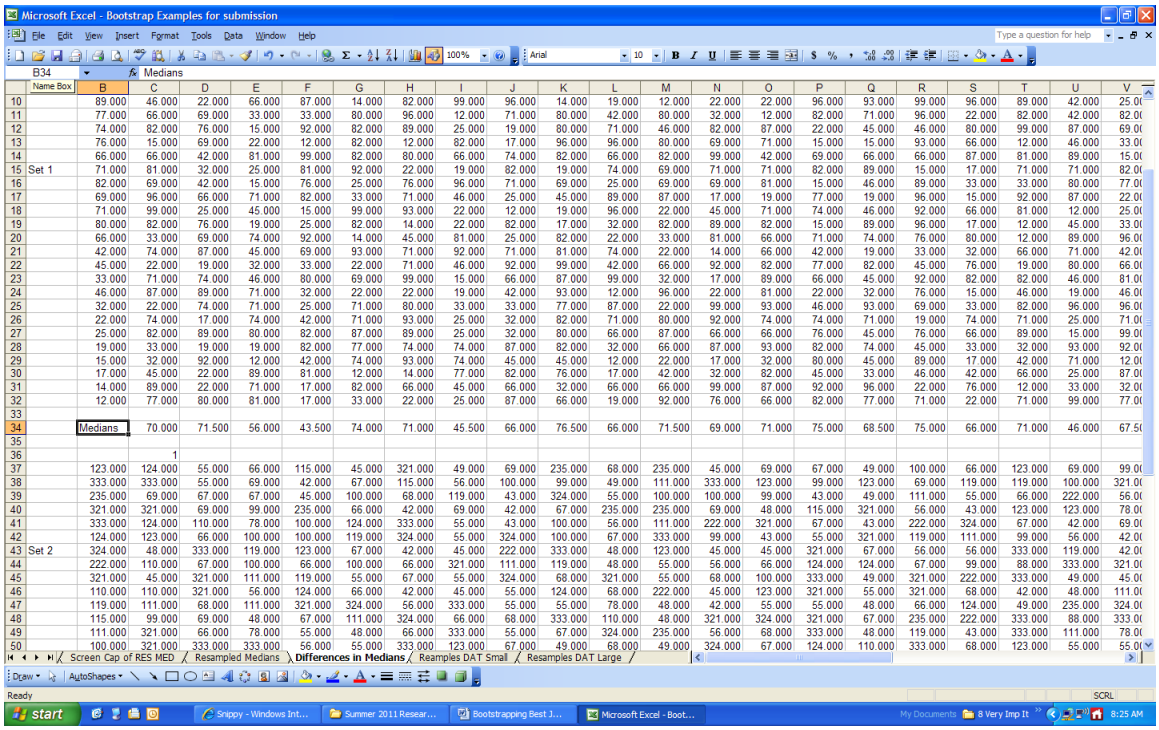

Figure 14: Display of Two Sets of Data

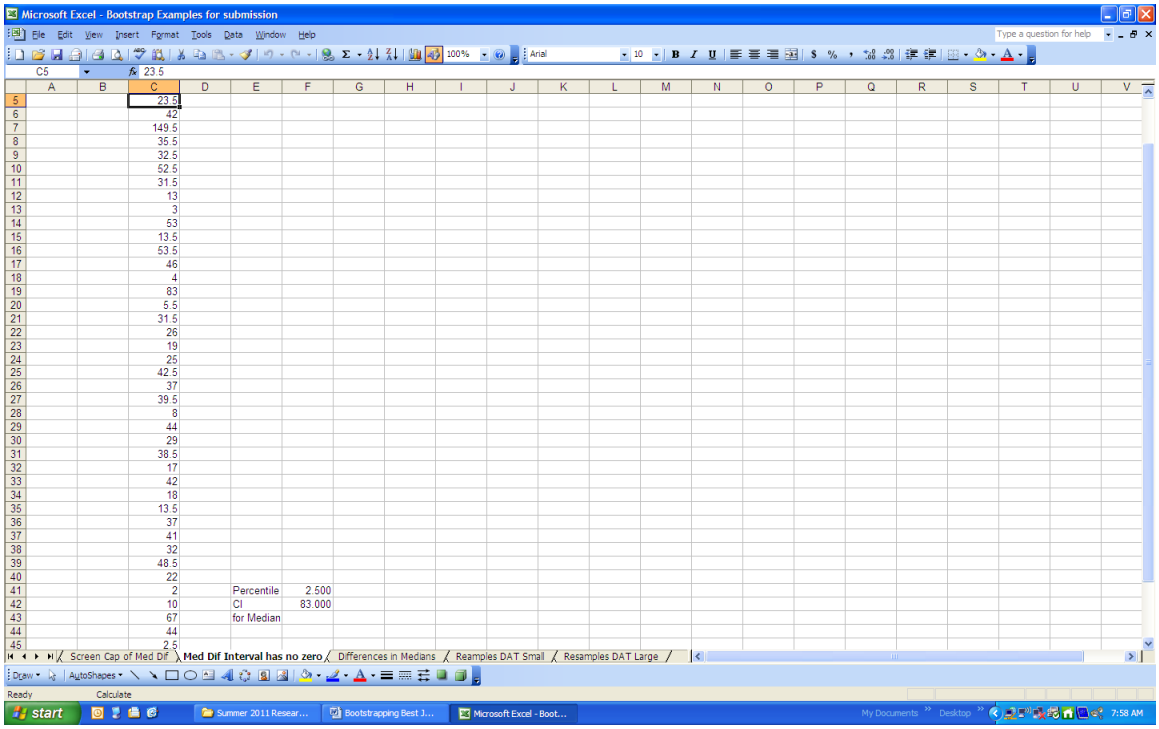

Figure 15: A Distribution of Median Differences with the Percentile Confidence Interval

The sampling distribution for the difference in medians is shown in figure 15. The 95% percentile confidence interval for the median difference is 2.5 to 83. Since 0 is not in the interval the null hypothesis of no differences in the median prices for the 2 groups is rejected. That is there is a difference in medians for the 2 sets of data. Note that pressing F9 many times provides about 95% of the intervals with 0 in them. This is in agreement with the concept of confidence interval. The medians of all these resamples were obtained by applying the function "=MEDIAN(range of cells)". For example "=MEDIAN(C9:P9)" calculates the median of the numbers in cells c9 through p9

6. Bootstrapping Analysis in EXCEL: Observations

Students as well as researchers learn in applying bootstrapping there is no need to satisfy assumptions when conducting inferential statistics. Also bootstrapping analysis can be performed on not only sampling distributions that are known but also on sampling distributions that are not known. Doing research this way enables the research community to recognize valid research findings even where assumptions are not justified.

Meaningful analyses can be made on populations that could not be analyzed before bootstrapping was developed. For example inferences can be made on population log means [5].

Results from the parametric and resampling approaches are comparable especially when the number of resamples is very large. With EXCEL spreadsheets, implementing a large number of calculations can be conducted.

EXCEL spreadsheets are valuable for bootstrapping analysis. Various capabilities of EXCEL for doing bootstrapping include the application of: a) the INDEX Function b) random numbers and c) the Data Analysis Toolpak and macros. As a result of using EXCEL, constructing confidence intervals, determining significance and graphing results or outcomes are easily done. There is almost no limitation on conducting tests of hypothesis for any statistic with bootstrapping.

7. Conclusions

EXCEL is easy to learn and useful for statistical analysis. Using the INDEX function, the Data Analysis Toolpak and automating calculations with macros provide students and researchers with a variety of useful techniques valuable not only for inferential statistics but for many other mathematical applications

There is more than one way to conduct valid statistics and all approaches lead to similar results.

Doing inferential statistics with bootstrapping resamples and without justifying assumptions allows a researcher to focus on results and not consider the ramifications and concerns of doing studies without verifying assumptions. The value of conducting bootstrapping analysis is in the ability to do valid statistical inferences without justification of any assumptions. Also where there is no known sampling distribution for a statistic, bootstrapping analysis can be performed.

The applications of confidence intervals for decision making in inferential statistics are valuable and can be constructed in EXCEL.

8. References

1. Devore, J.L. *Probability and Statistics for Engineering and the Sciences 7th Edition.* California: Thomson Brooks/Cole.
2. Diaconis, P., and Efron, B. (1983). *Computer-intensive methods in statistics*. Scientific American, May, 116-130.
3. Hesterberg, T., Monaghan, S., Moore, D. S., Clipson, A, and Epstein, R. (2003). *Bootstrap Methods and Permutation Tests*. Companion Chapter 18 to The Practice of Statistics. New York: WH Freeman and Company.
4. Mendenhall, W., Beaver, RJ and Beaver, BM (2006). *An Introduction to Probability and Statistics. 12th edition*. California: Thomson Brooks/Cole.

5. Mooney, C.Z. and Duval, R. D. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference*. California: Sage.
6. Moore, D., McCabe, S, George P and Bruce, C. (2009). *Introduction to the Practice of Statistics 6th edition*. New York: WH Freeman Co.
7. Ott, R L. and Longnecker, M. (2001). *An Introduction to Statistical Methods and Data Analysis 5th Edition*. California: Thomson Duxbury.

9. Appendices

9.1: Installing Data Analysis Toolpak

Before using Excel install the Data Analysis Toolpak for MS Office 2003 by going into EXCEL under Tools>Options> Add-ins and checking off Data Analysis Toolpak.

For MS Office 2007, select the Microsoft Office button and choose EXCEL Options>Add-ins>Manage>choose EXCEL Add-ins>Go. When in the add-ins box>choose Analysis Toolpak>ok

In MS Office 2003 or MS Office 2007, after installation, Data Analysis is available on the EXCEL DATA tab. If the user wishes to use Visual Basic Applications (programming capabilities) selecting Data Analysis Toolpak VBA is also available.

With the Data Analysis Toolpak a variety of statistical techniques including the ability to create frequency tables and histograms, determine descriptive statistics and perform various inferential statistics are available.

9.2: Insert frequency distributions and histograms using Data Analysis Toolpak.

a) Select Histogram
b) Check off Chart
c) Select Bin, these numbers can be obtained by finding the largest and smallest number in the data.
d) Run the command
e) Select or highlight input and highlight where to display output

9.3: Bootstrapping With Data Analysis Toolpak

Resampling with replacement can be done easily in Excel. Here is how:

1) Install add-ins Data Analysis Toolpak and Data Analysis Toolpak VBA.
2) Under tools select data analysis.
3) Select sampling.

4) Highlight the data you wish to resample
5) Select a range where you want resampled data to be placed.
6) Select the number of samples you wish
7) Click on enter and your resampled data is displayed where requested

See example 3 and figure 9 for a display of results.

9.4: Macro for resampling

The following macro resamples the given dataset 40 times:

```
Sub Test()
Range("D7:Aq36").Select
Selection.ClearContents
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$D$7:$d$36"), "R", 30, False
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$e$7:$e$36"), "R", 30, False
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$f$7:$f$36"), "R", 30, False
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$g$7:$g$36"), "R", 30, False
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$h$7:$h$36"), "R", 30, False
```

```
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$am$7:$am$36"), "R", 30, False
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$an$7:$an$36"), "R", 30, False
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$ao$7:$ao$36"), "R", 30, False
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$ap$7:$ap$36"), "R", 30, False
Application.Run "ATPVBAEN.XLA!Sample",
ActiveSheet.Range("$c$7:$C$36"),
ActiveSheet.Range("$aq$7:$aq$36"), "R", 30, False
```

```
'8 groups of 5 show 40 resamples
End Sub
```

9.4.1: Discussion about macro

1) Line 1 selects a range of cells then line 2 clears contents.

The macro is documented by using apostrophe as in the above '8 groups of 5 show 40 resamples".

The code segment (`Application.Run "ATPVBAEN.XLA!Sample", ActiveSheet.Range("$c$7:$C$36"),ActiveSheet.Range("$D$7:$d$36"), "R", 30, False)` does the actual resampling once. In order to resample more than one time change the segment ("$C$7:$C$36") to any desired location (such as ("aq7:aq36") for resampling given data and displaying more than 1 resample. The code displayed finds resamples of size 30

2) Typical output using this macro for 40 resamples of size 30 is presented in figure 11
3) The process the user goes through to accomplish the above steps can be done by recording a macro.
4) ATPVBAEN (Analysis Toolpak for Visual Basic Applications in English) is set of programs in VBA necessary for running macros. When in EXCEL using Tools> Add-in check off Analysis Toolpak and Analysis Toolpak VBA, ATPVBAEN is part of the add-in Analysis Toolpak.

Using the Toolpak and EXCEL's charting capability, a frequency distribution as well as a histogram can be displayed. Figure 11 is a screen capture of the output of the frequency distribution and histogram. Notice the frequency distribution and the shape of the histogram data appears to be mound-shaped.

Using Toolpak is invaluable for obtaining frequency distributions and histograms. In order to get a better visualization of the distributions developed, a bar chart can be inserted once data are ordered in a frequency table as shown in the above screen capture. The bar chart for resampled means is also illustrated in figure 11.