THE IMPORTANCE OF TEACHING POWER IN STATISTICAL HYPOTHESIS TESTING

Alan Olinsky, Bryant University, (401) 232-6266, aolinsky@bryant.edu *

Phyllis Schumacher, Bryant University, (401) 232-6328, pschumac@bryant.edu

John Quinn, Bryant University, (401) 232-6392, jquinn@bryant.edu

KEY WORDS: statistical power, statistics education, hypothesis testing

*Corresponding author

1 An earlier, preliminary form of this paper was presented at the Northeast Decision Sciences Annual Meeting, 3/28-3/30/2008 and was published in the Proceedings
ABSTRACT

In this paper, we discuss the importance of teaching power considerations in statistical hypothesis testing. Statistical power analysis determines the ability of a study to detect a meaningful effect size, where the effect size is the difference between the hypothesized value of the population parameter under the null hypothesis and the true value when the null hypothesis turns out to be false. Although power is an important concept, since not rejecting a false hypothesis may result in serious consequences, it is a topic not often covered in any depth in a basic statistics class and it is often ignored by practitioners. Considerations of power help to determine appropriate sample sizes for studies and also force one to consider different effect sizes. These are important concepts but difficult for beginning statistics students to understand. We illustrate how one can provide a simple classroom demonstration; using applets provided by Visual Statistics 2.0, of how to calculate power and at the same time convince students of the importance of power considerations. Specifically, for beginning students, we focus on a common statistical hypothesis testing example, a test of hypothesis of means concerning one sample. For this case, we examine the power of the test at varying levels of significance, sample sizes, standard deviations and effect sizes, all factors which are important to the results of a testing situation. We then illustrate how students, depending on time and resources, can reproduce these power calculations themselves using several statistical software packages. The use of statistical software will also be very helpful to the students who upon graduation become the practitioners. Finally, examples of power analyses are provided for more advanced problems such as analysis of variance and regression analysis, which might be used in a second semester statistics course.
INTRODUCTION

In the statistical application of hypothesis testing, the significance level, or p-value, is always emphasized but often the power of the test is not given as much importance as it should have. The significance level is the probability of rejecting a true null hypothesis, also described as committing a Type I error. Hypothesis tests are often called significance tests and when performing a hypothesis test, one is encouraged to reject the null hypothesis if the evidence that it is incorrect is overwhelming. In the absence of statistical significance, rather than accepting that the null hypothesis is true, it is commonplace to conclude that it cannot be rejected thus avoiding committing a type II error, which is the error of accepting the null hypothesis when the alternative hypothesis is true. Indeed the common analogy of comparing hypothesis testing to the U.S. legal system, which assumes that the accused is innocent beyond a reasonable doubt or acquitted in the absence of enough evidence, encourages the consideration of the level of significance to the exclusion of considerations of the probability that the alternative hypothesis might be true. Many one-semester elementary statistics texts either briefly define power with reference to Type II error without providing examples of how to calculate power or they completely ignore the topic.

In most two semester basic texts, examples and problems are provided but students find the topic very difficult. Therefore, since students find understanding power difficult, it is convenient to limit testing to consideration of levels of type I error rather than deal with power or increases in sample size beyond a brief definition or perhaps a table of power considerations called an Operating Characteristic Curve. However, with large samples it is easy to reach statistical significance. A statistical difference is not the same as a practical difference. With small samples, practical differences may not be statistically significant and vice-versa, with very large samples statistical differences may not be of practical importance.

The importance of power in real applications of statistical hypothesis is well documented. Many authors in diverse fields have addressed the importance of these considerations. Several examples are given below. Thomas and Juanes (1996) indicate that in biology, for a large enough sample size “any statistical hypothesis test is likely to be statistically significant, almost regardless of the biological importance of the results.” Bhardwaj et al. (2004) discuss the importance of power considerations in clinical trials. They caution that: “Dermatologists should not focus on small p-values alone to decide whether a treatment is clinically useful; it is essential to consider the magnitude of treatment differences and the power of the study.” They go on to explain that the “relationship between sample size and power is critical to interpreting the conclusions that can be drawn from a study.” Yaffee (1997) agrees on the importance of proper preliminary power and sample size analysis and states that “If medical researchers are engaged in clinical trails of a drug, insufficient sample size that undermines assessment is criminal” Chester

2 An earlier, preliminary form of this paper was presented at the Northeast Decision Sciences Annual Meeting, 3/28-3/30/2008 and was published in the Proceedings
Spatt, the Chief Economist for the Office of Economic Analysis, in a recent memorandum to the Investment Company Governance, highlights the importance of power considerations and sample size when studying mutual fund returns (Spatt, 2006). He reports that according to their analysis, “most studies assessing the impact of chair independence on returns do not have sufficient power to reliably conclude that a relationship does or does not exist.” Heping Deng (2005), in a study of educational research, warns that the economic cost of the implementation of a new program, assumed to have improved results, will be wasted if insufficient evidence is obtained to determine true gains. He concludes that: “a powerful significance test can lead to better decision making by educational leaders.” A final and interesting example is provided by a report found on the New York Department of Health website concerning power and the fears aroused in the Love Canal health study. The article written discusses some of the limitations of the study caused by the low power to detect small differences in more common health effects, even though the statistics used have strong power in detecting rare illnesses. They conclude that “Not seeing a difference, especially when the power is low, does not mean there is no difference and reporting no difference could be misleading and cause someone not to be concerned about a real health effect.” The committee in fact concludes that “we should not rely on statistical significance as a way to determine biologic import.” (NY Dept. of Health)

Therefore, even though it is more elusive, it is important not to ignore the possible alternative hypotheses and effect sizes (differences between the hypothesized and actual value of a parameter) and the calculation of the power of a test, that is, the probability of rejecting a false null hypothesis. These considerations are important both for practitioners, as indicated by the studies cited above, and for statistics students at all levels. Students should always be aware of the consequences of not rejecting hypothesis when they are false. In order to illustrate the importance of power and how it relates to type I error, sample size, and effect size, classroom demonstrations are helpful, particularly for beginning students.

Using appropriate statistical software packages and relevant examples, we illustrate the computation of power for hypothesis testing. Specifically, we use Visual Statistics 2.0 (Doane et al., 2001) for demonstration purposes and SAS and MINITAB, both commonly used in basic statistics classes, to compare and contrast the ease of calculating power and/or sample size effects (SAS, 2004; MINITAB, 2007).

Many texts provide formulas for calculations of sample size needed to achieve a given significance for varying margins of error when performing hypothesis tests but these often ignore power. In a more advanced course, for example a course including multiple regression, guidelines are often given as to how many cases are needed for acceptable levels of power, but these are typically general in nature and do not provide any insight into how to calculate power. Gatti and Harwell (1998), discuss the use of computer programs for estimating power rather than referring to power charts. They point out in their paper that many statistics and research design textbooks do highlight the importance of power considerations for empirical studies. They urge that students at this level be taught to estimate power using statistical programs such as SPSS and SAS. We propose a combination of demonstration and simple calculations that can be used to teach beginning statistics students about power and its importance. This analysis will generate sample sizes necessary under different scenarios to achieve an acceptable power level, which is often recommended to be a minimum of 0.8 (Hair, 2006).
It is important to remember that there are limitations to the use of power. One caveat to keep in mind is that pointed out by Pinto and Allen (2003) with regard to the significance level of the test. As indicated in their paper, power is only important if the hypothesis is false and is in fact not defined and should not be used to describe cases where the null hypothesis is true. A second limitation of power, referred to by Lenth (2001) is that it is important to note that a power analysis should be carried out prospectively rather than retrospectively. As he points out, “using retrospective power for making an inference is a convoluted path to follow. The main source of confusion is that it tends to be used to add interpretation to a non-significant statistical test.” This is also emphasized by Goodman and Berlin (1994) who observe that, with medical research and clinical studies; “power should play no role once data have been collected but exactly the opposite is widely practiced.”

The initial portion of our analysis will focus on a common statistical hypothesis testing example, that is, a test of hypothesis of a population mean, first with a classroom demonstration and then by illustrating how students can produce these results themselves with a statistical package. We will examine the power of the test at varying levels of significance, sample size, effect size, and standard deviations. We present our results in tabular and graphical format.

There are examples of power being used with more advanced statistical procedures. This analysis can be done in SAS and other statistical packages. A good online reference for power analysis for these more advanced analyses can be found at the UCLA Academic Technology home page. We will provide examples from a popular business statistics text which illustrate how this power analysis can be performed by the students using SAS. A detailed explanation of these SAS procedures can be found at the UCLA Academic Technology page for which the address is provided in the bibliography. (Introduction to SAS, 2010).

CLASSROOM DEMONSTRATION

A simple way to explain power and its importance is by using a classroom demonstration. These demonstrations can be performed with applets which are available online or included in software such as Visual Statistics. We will provide an illustration of such a demonstration using a simple example and illustrations from Visual Statistics 2.0 (Doane et al., 2001). Power curves are presented which illustrate how the power is affected by changes in the sample size, changes in levels of significance and at different levels of the effect size. The graphical output is a very attractive aspect of using this program. The example we have selected is the same one used by Pinto and Allen (2003) when they caution against the use of power for cases when the null hypothesis is true.

The voltage of a new battery is supposed to be 1.5 volts. The population standard deviation is believed to be 0.1 volts. A random sample of 36 new batteries is taken and the voltage of each battery is measured. A two-tailed test of the null hypothesis that the population mean is 1.5 is to be performed. The power of this test is then examined in the case where it turns out that the true mean is actually 1.45 volts (an effect size of .05) is then analyzed by varying three quantities: the sample size, the level of significance, and the standard deviation.
The Visual Statistics output for this example is presented in Figure 1. The illustration actually represents a family of power curves which plot the power as the effect size changes for varying sample sizes for the case of a null hypothesis of $\mu=1.5$, with a two tailed test at a level of significance of 0.01 and a standard deviation of 0.1. The vertical axis represents power and the horizontal axis possible values of the population mean.

![Figure 1: Family of Power Curves Varying Sample Size (H₀: $\mu = 1.5$, H₁: $\mu \neq 1.5$, True Mean = 1.45, $\alpha = 0.01$, $\sigma = 0.1$, n = 18, 36, 72)](image)

One can see, as indicated with dotted lines in the middle curve, with n=36, when the alternative is 1.45, the power is 0.66. The top curve, with an increase in sample size to 72, shows that at the same alternative the power increases to almost 1. With a decrease in sample size to 18, illustrated in the bottom curve, the power decreases to approximately 0.3. One can also observe that if the effect size is zero, that is the null hypothesis is true ($\mu=1.5$), the power curve indicates a value of 0.01 which is the level of significance of the test and is not accurately “power” as indicated by Pinto. One can also see in this picture that as the effect size increases, in both directions, the power also increases for all sample sizes.

Since in general, for a fixed sample size with a specific alternative hypothesis, an increase in the probability of Type I error will result in a less significant result. This increase will also cause a
decrease in the probability of Type II error and therefore an increase in power. So, if one fixes the sample size and the standard deviation and varies the level of significance, as in done in Figure 2, one should see that as the level of significance (Type I error) increases, the power will also increase. In Figure 2, with the sample size fixed at 36 and the standard deviation assumed to be 0.1, we look at levels of significance of 0.01 (bottom curve), 0.05 (middle curve), and 0.1 (top curve). It can be seen that, as expected, as Type I error increases the power increases at a specific effect size, with again larger power at larger effect sizes.

Finally, we investigate how changes in variation affect power. In most testing situations where the true population mean is unknown, the true value of the variance is also unknown but often a good estimate can be obtained. This is sometimes done by previous tests or by pre-sampling but most often the sample variance is used. For increases in variation, there will be increases in error and so the power will decrease. Figure 3 illustrates how the power changes with changes in the standard deviation. Here, the level of significance of 0.01 and the sample size of 36 are kept constant and the power curves are calculated for standard deviations of 0.05 (top curve), 0.1 (middle curve) and 0.2 (bottom curve). In this case, we confirm that for a given effect size, as the standard deviation increases the power does indeed decrease.
Figure 3: Family of Power Curves Varying Standard Deviation ($H_0: \mu = 1.5$, $H_1: \mu \neq 1.5$, True Mean = 1.45, $n = 36$, $\alpha = 0.01$, $\sigma = 0.05$ (light green – top curve), 0.10 (light blue – middle curve), 0.20 (dark blue – bottom curve))
(Note: Graphical output from Visual Statistics has incorrect legend for standard deviations)

**COMPUTATIONS OF POWER USING SAS**

Visual Statistics has provided an insightful visual demonstration of power and how it responds to changes in effect size, sample size, level of significance and variation. This is a useful tool for classroom use. Popular statistical packages such as SAS and MINITAB, which are commonly used in both statistics courses and the real world, can also be utilized to demonstrate some of the same points. Perhaps an even more advantageous pedagogical consideration is that the students themselves can do the calculations to determine appropriate sample sizes for obtaining powerful hypothesis tests in particular applications. The advantage of using one of these statistical programs is that the students then have a tool that they can utilize in future classes or in the workplace.

The changes in power obtained by Visual Statistics were illustrated in Figures 1-3. Using SAS to illustrate the same example, the effect on power of variations in effect size, sample size, and level of significance can all be combined in one procedure for a given standard deviation. An interactive version of SAS was used to obtain SAS output but the results can be obtained using
other versions of SAS with the Proc Power command. SAS commands are provided in the examples for more advanced power analyses which are provided further on in this paper.

The output from SAS is provided in Table 1 for a null mean of 1.5 and a standard deviation of 0.1 and a two-sided alternative. Power calculations are obtained for varying effect sizes (alternate means 1.45 and 1.49) at different levels of significance (0.01, 0.05, and 0.1) and for various sample sizes from 18 to 72. With an alternate mean of 1.45, it is also interesting to note that in order to reach the power of 0.8, with a significance level of 0.01, approximately 54 observations are needed, at a significance level of 0.05, only 36 observations are needed and at a significance level of 0.1, fewer than 30 observations.

### Table 1: SAS Output of Power Varying Effect Size and Level of Significance

<table>
<thead>
<tr>
<th>Alternate Mean</th>
<th>Alpha</th>
<th>N</th>
<th>Power</th>
<th>Alternate Mean</th>
<th>Alpha</th>
<th>N</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.45</td>
<td>0.01</td>
<td>18</td>
<td>0.255</td>
<td>1.49</td>
<td>0.01</td>
<td>18</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36</td>
<td>0.610</td>
<td></td>
<td></td>
<td>36</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td></td>
<td>54</td>
<td>0.837</td>
<td></td>
<td></td>
<td>54</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72</td>
<td>0.941</td>
<td></td>
<td></td>
<td>72</td>
<td>0.040</td>
</tr>
<tr>
<td>1.45</td>
<td>0.05</td>
<td>18</td>
<td>0.516</td>
<td>1.49</td>
<td>0.05</td>
<td>18</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36</td>
<td>0.830</td>
<td></td>
<td></td>
<td>36</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td></td>
<td>54</td>
<td>0.950</td>
<td></td>
<td></td>
<td>54</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72</td>
<td>0.986</td>
<td></td>
<td></td>
<td>72</td>
<td>0.133</td>
</tr>
<tr>
<td>1.45</td>
<td>0.10</td>
<td>18</td>
<td>0.652</td>
<td>1.49</td>
<td>0.10</td>
<td>18</td>
<td>0.128</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36</td>
<td>0.902</td>
<td></td>
<td></td>
<td>36</td>
<td>0.158</td>
</tr>
<tr>
<td></td>
<td></td>
<td>54</td>
<td>0.976</td>
<td></td>
<td></td>
<td>54</td>
<td>0.187</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72</td>
<td>&gt;.99</td>
<td></td>
<td></td>
<td>72</td>
<td>0.217</td>
</tr>
</tbody>
</table>

### One-Sample t-Test

Null Mean = 1.5  Standard Deviation = .1  2-Sided Test

SAS output can also be generated with JMP (2007) which is an interactive visual package for statistical analysis produced by SAS. JMP is becoming more and more popular in statistics classes. The output is similar to what was obtained with Visual Statistics and it does produce power for tests of means.

**COMPUTATIONS OF POWER USING MINITAB**

MINITAB is a statistical software package that is widely used in basic statistics courses and also by practitioners. It can be utilized to provide the same results as previously obtained by Visual Statistics and SAS. In the MINITAB illustration, the variance is fixed at 0.1, the level of
significance is fixed at 0.05 and the power is requested for an effect size of 0.5 and samples sizes of 18, 36, 54, and 72. The output will contain both tables and graphs, the graph will be provided below. As is true with SAS, the students can obtain the output themselves and then utilize the procedure on other problems. MINITAB is a menu driven interface. The menu choices necessary to obtain the output are as follows: Stat->Power and Sample Size->1-sample t->. These choices produce the table in Figure 4, where one inputs the sample sizes and effect size in order to produce the corresponding power.

![MINITAB Menu for Power Calculations for a One Sample Test for a Population Mean](image)

**Figure 4: MINITAB Menu for Power Calculations for a One Sample Test for a Population Mean**

The MINITAB output is provided in Figure 5 which illustrates changes in power as the effect size and sample size are varied. This output is essentially the same as the Visual Statistics output provided in Figure 1. It should be noted that both SAS and MINITAB use a t-distribution to obtain power calculations whereas Visual Statistics uses the standard normal distribution and the power calculations are, therefore, slightly different. As indicated earlier, in most cases when testing for a population mean the population variance is not known in which case the t-distribution is the preferred statistic and this is particularly important for small samples where the standard normal may not provide a good estimate of the t-statistic.
COMPUTATIONS OF POWER FOR ANALYSIS OF VARIANCE

As mentioned in the introduction, power calculations are more difficult to understand in more advanced statistical estimation procedures such as Analysis of Variance (ANOVA) or Regression. We will now provide an analysis using SAS which computes power for an example where the appropriate test is a One-way ANOVA.

For this illustration, we will use an example from the text *Statistics for Business and Economics* by McClave, Benson, and Sincich (2008), which describes an ANOVA F-test to compare golf ball brands but does not include any power calculations. The example presents data used to compare the means the distance travelled of four different brands of golf balls when hit with a driver. A completely randomized experiment is utilized by having a robotic golfer hit 10 golf balls of four different brands of golf balls using a driver. The four sample means are: 250.8, 261.1, 270.0, and 249.3. The average sample standard deviation for these four samples was 4.6. The ANOVA is run with a significance level of 0.1. This example results in an observed F value of 43.99 and a p-value of 3.97E-12, clearly very significant results.

We now illustrate how students can perform a power analysis for this example using SAS. The output in Figure 6 includes the SAS commands and a portion of the SAS output which has been edited to illustrate several scenarios. It should be noted that when running the Proc Power for ANOVA, with fixed scenario elements, one inputs alpha, the standard deviation, and the group means and one can omit either the power by entering a period and include the sample size or omit the sample size by entering a period and proposing a value for the power. The commands and full output is included for the first scenario but for the cases II and III where all fixed elements remain the same except the group means, only the group means and corresponding power are included. In Scenario IV, V and VI, the Method, Alpha, and standard deviation is kept the same but instead of assuming that the samples have 10 outcomes each, the procedure is run to
find the sample size needed to obtain a power of at least 0.8. The output obtained when one runs the power procedure with a level of significance of 0.1, a standard deviation of 4.6 and the sample means as stated in the above example is shown below in Figure 6. It can be seen that the Power for this test is very high. This is not surprising since the sample means are quite different and we are finding the probability of finding populations to not all be the same in the case that they are as presented. One can see that if the Power Procedure is rerun with different means which are closer together, a test with all other elements kept the same is less powerful.

```plaintext
proc power ;
onewayanova
  groupmeans = 250.8 | 261.1 | 270 | 249.3
  stddev = 4.6
  alpha = 0.10
  npergroup = 10
  power = .;
run;
```

The POWER Procedure
Overall F Test for One-Way ANOVA
Fixed Scenario Elements

<table>
<thead>
<tr>
<th>Method</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>0.1</td>
</tr>
<tr>
<td>Group Means</td>
<td>250.8 261.1 270 249.3</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>4.6</td>
</tr>
<tr>
<td>Sample Size Per Group</td>
<td>10</td>
</tr>
</tbody>
</table>

Computed Power - Power>.999

Scenario II
Group Means 260 263 265 268
Computed Power Power = 0.952

Scenario III
Group Means 260 261 262 263
Computed Power Power = 0.312

Scenario IV
Group Means 260 261 262 263
Nominal Power 0.8
Computed N Per Group

<table>
<thead>
<tr>
<th>Actual N Per Power Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.810</td>
</tr>
<tr>
<td>39</td>
</tr>
</tbody>
</table>

Scenario V
Group Means 260 263 265 268
Nominal Power 0.8
Computed N Per Group

<table>
<thead>
<tr>
<th>Actual N Per Power Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.843</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>

Scenario VI
Group Means 250.8 261.1 270 249.3
**Figure 6: SAS Commands and Output for Power Analysis of an ANOVA F-test.**

### COMPUTATIONS OF POWER FOR REGRESSION

The importance of power considerations in regression is noted by Eisenhauer (2009). He points out that “very little explanatory power is required in order for regressions to exhibit statistical significance.” Using software, students can investigate power with regard to a problem in multiple regression. We will use SAS and an exercise from the section on comparing nested regression models in the McClave textbook. The following example involves obtaining the best model for predicting the dependent variable \( y = \) heat rate (kilojoules per kilowatt hour) of a gas turbine as a function of two variables \( X_1 = \) cycle speed (revolutions per minute) and \( X_2 = \) cycle pressure ratio. (McClave, p.733) The data set for the exercise has 67 observations. The exercise involves fitting two models. The first model is a complete second order model which would predict the dependent variable, \( y \) five independent variables, \( X_1, X_2, \) the interaction term \( X_1 \times X_2 \) and two quadratic terms \( X_1^{Sq} \) and \( X_2^{Sq} \). This model is run and is found to fit well with an R-square of 88.5%. The second model is a reduced model which omits the quadratic terms and keeps the other three terms. This model is run and is also found to fit well with an R-square of 84.9%. The question in the textbook exercise is whether or not the quadratic model provides a significant improvement over the reduced model thereby justifying the use of a more complicated model. The test for the comparison is a partial F test. Since an increase in the number always increases the value of R-square, the real question asked concerns the size of the increase. For the purposes of this paper, we suggest that the student can investigate, what sample size would be necessary to detect a difference in R-square of 0.05, that is an increase from 84.9% for 3 predictor variables to 88.5% with 5 predictor variables, that is when testing the increasing in R-square due to the 2 assessed variables, which are called test predictors in the SAS commands. The output for this analysis appears below in Figure 7. The SAS commands appear in the top of the output. Note that the procedure has been run in order to find sample size. This is indicated by the statement “ntotal=.”.

It can be seen in Figure 7 that for a regression equation with 3 independent variables and an \( R^2 \) of 0.849, if one were to add 2 more independent variables, a sample size of 29 is needed to increase the value of \( R^2 \) by 5% to 0.885 and obtain a power of 0.7. To obtain a power of .8 with \( R^2=0.885 \) a sample size of 35 is required and for a power of 0.99 the sample size would need to be 44.

```sas
proc power;
   multreg
      model=fixed
      nfullpredictors = 5
      ntestpredictors = 2
      rsquarefull=.885
      rsquarediff=.036
```

---

**Nominal Power**

0.8

**Computed N Per Group**

<table>
<thead>
<tr>
<th>Actual N Per Group</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.898</td>
<td>2</td>
</tr>
</tbody>
</table>
CONCLUSION

In summary, practitioners and students of all levels should be aware of the importance of considerations of statistical power when conducting hypothesis tests. This analysis allows one to determine the ability of a study to detect a meaningful effect size. As we have shown above students in elementary statistics classes can be taught by simple demonstrations how to determine sample sizes necessary to achieve appropriate power levels (at least 0.8 according to Hair (2006)). The demonstration also shows that in order to achieve the desired power level, more stringent significance levels (e.g., .01 instead of .05) and/or smaller effect sizes require larger samples, and smaller standard deviations increase the power of a test procedure.

If time and resources permit students can also use the MINITAB and SAS software procedures which to perform power analyses themselves. These statistical packages are menu driven and easy to use. For a more advanced student, power analysis can be performed for more sophisticated testing procedures. Another source of information about performing power analysis with software is a paper by Robert Yaffee (1997), which provides a comparison evaluation of several packages designed specifically for power analysis. We provided examples for Analysis of Variance and Multiple Regression which are done in SAS along with a reference to find more detailed discussion online (Introduction to SAS, 2010). In addition to this reference, there are several more interesting examples of advanced power calculations and additional software

<table>
<thead>
<tr>
<th>Fixed Scenario Elements</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Exact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Fixed X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Predictors in Full Model</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Test Predictors</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-square of Full Model</td>
<td>0.885</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference in R-square</td>
<td>0.036</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alpha</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Computed N Total</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
<td>Nominal Power</td>
<td>Actual Power</td>
<td>N Total</td>
</tr>
<tr>
<td>1</td>
<td>0.7</td>
<td>0.715</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>0.810</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>0.901</td>
<td>44</td>
</tr>
</tbody>
</table>

Figure 7: SAS Output for Power Analysis of the Comparison of Two Multiple Regression Models.
package power procedures as illustrated by Park (2008) and also by Gatti and Harwell (1998). There are even free calculator applications available on the Web to perform power analysis with different models. One such website is [www.danielsoper.com](http://www.danielsoper.com).

As indicated in the references cited considerations of power are very important. This is particularly true in medical studies such as the Love Canal study. Although it is not as straightforward a concept as the significance level of a test, power analysis can and should be a part of every statistics course. We have shown that, through easy to produce software demonstrations, the importance of power can be conveyed to students, even in basic statistics classes. We have also demonstrated that power calculations are easily obtainable through the use of many popular statistical packages. Thus, when doing hypothesis testing and planning research studies, it is important to consider power when determining appropriate sample sizes. Such proper planning will reduce the risk of conducting a study that doesn’t produce useful results. For more information and advice on teaching power analysis, we recommend Christopher Aberson’s “An Interactive Tutorial for Teaching Statistical Power,”(Aberson et al, 2002)
REFERENCES


