

Journal of Statistics Education



ISSN: (Print) 1069-1898 (Online) Journal homepage: https://www.tandfonline.com/loi/ujse20

Making the Concepts of Power and Sample Size Relevant and Accessible to Students in Introductory Statistics Courses using Applets

C. M. Anderson-Cook & Sundar Dorai-Raj

To cite this article: C. M. Anderson-Cook & Sundar Dorai-Raj (2003) Making the Concepts of Power and Sample Size Relevant and Accessible to Students in Introductory Statistics Courses using Applets, Journal of Statistics Education, 11:3, , DOI: 10.1080/10691898.2003.11910721

To link to this article: https://doi.org/10.1080/10691898.2003.11910721





Making the Concepts of Power and Sample Size Relevant and Accessible to Students in Introductory Statistics Courses using Applets

C. M. Anderson-Cook and Sundar Dorai-Raj

Journal of Statistics Education Volume 11, Number 3 (2003), ww2.amstat.org/publications/jse/v11n3/anderson-cook.html

Copyright © 2003 by C. M. Anderson-Cook and Sundar Dorai-Raj, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Internet; Java applet; One-sample hypothesis test; Two-sample hypothesis test.

Abstract

The concepts of hypothesis testing, trade-offs between Type I and Type II error, and the use of power in choosing an appropriate sample size based on power when designing an experiment are routinely included in many introductory statistics courses. However, many students do not fully grasp the importance of these ideas and are unable to implement them in any meaningful way at the conclusion of the course. This paper presents a number of applets intended to help students understand the role of power in hypothesis testing and which allow them to obtain numerical values without having to perform any calculations for a variety of scenarios, complementing some of the applets presented in Aberson, Berger, Healy, and Romero (2002). Ideas are given about how to incorporate the materials into an introductory course.

1. Introduction

As noted in Aberson, Berger, Healy, and Romero (2002), hypothesis testing is a conceptually rich topic in statistics that is frequently cited as one of the most difficult areas for students in introductory classes to master. The choice of test, identification of the correct calculations, interpretation of a critical value or a p-value are some of the complications that make for a difficult topic. Conceptually, students need to grasp that hypothesis testing is based on the idea of proof by contradiction. We assume that the null hypothesis is true and look for evidence in the data to either support or reject the null hypothesis, H_0 . If we reject the null hypothesis, we are concluding that the alternative hypothesis or research hypothesis, H_0 , is more plausible. A fundamental concept of the method is the notion that we will continue to assume that the null hypothesis is true until there is overwhelming evidence against H_0 (typically, less than a 1% or 5% chance of obtaining the observed value or one more extreme if in fact the null hypothesis were true).

Once students have learned how to perform a test (either with hand calculations or software), little attention is typically spent exploring what happens with the test when the alternative hypothesis is correct. The concept of power is frequently a lightly covered topic in introductory classes. The major reasons for this are the complication of the many scenarios to be considered and the difficulty of the associated calculations.

While the light coverage of the notion of power is understandable, given student problems with grasping the basic hypothesis testing concepts, and time constraints, it does represent a significant omission in students' statistical education. As statistical consultants who help researchers in a wide variety of disciplines with experiment planning and design, we commonly see experimenters completely ignore this aspect of hypothesis testing in the design phase. This can lead to any of several problems. First, they may use a sample size considerably larger than needed, wasting valuable resources. More commonly, they can use a sample size that is too small, so that only an unreasonably large effect could be detected. For a researcher, deciding what size of difference between treatment groups will give a meaningful result is crucial to the success of the study. This assessment should typically be performed before any data is collected, in order that the results of the data do not cloud his or her judgment, and to avoid wasting resources. If students do not have an adequate grasp of power, then these issues are ignored. Cohen (1988) also noted some of these issues and misconceptions.

Several excellent websites exist that offer different options for demonstrating or calculating power. The WISE site (http://wise.cgu.edu) referenced in Aberson et al. (2002)) offers a teaching tool for showing the connection between the distribution of the original observations, the distribution of the sample average, and power, all in the context of the one sample test for the mean. It allows flexibility to change many parameters and observe the effect, graphically and numerically, on the test. Excellent teaching resources for how to use the material are provided online and in Aberson et al. (2002). Other good sites for illustrating power are Todd Odgen's site

(http://www.stat.sc.edu/~ogden/javahtml/power/power.html) and Java Applets for Visualization of Statistical Concepts from Katholieke Universiteit Leuven (http://www.kuleuven.ac.be/ucs/java), which both provide primarily graphical visualization aids for understanding power. A very basic site with a Flash demo for helping students master the terminology of Type I error and Type II error, and power can be found at (http://www.psych.utah.edu/learn/statsampler.html#Powertool). For calculation of power, the most comprehensive site we have found is by Russ Lenth

(http://www.stat.uiowa.edu/~rlenth/Power), which allows flexible computation with some limited graphical summaries of many different scenarios of hypothesis testing. However, we feel this site would most sensibly be used by a knowledgeable statistician, rather than as a teaching aid for introductory statistics classes.

In this paper, we present applets that demonstrate and quantify the effects of changing the sample size and alpha-level on the power for the one- and two-sample tests for both means and proportions. Gatti and Harwell (1998) stress the advantages of using computers to compute power rather than hand-calculations or tables, and these applets provide a means for calculations as well as visual representation of concepts behind them. These applets are freely available on the internet, and have been successfully implemented into a number of statistics courses at Virginia Tech. Nickerson (1995) outlines five principles to foster understanding, which include "start where the student is", "promote active processing and discovery" and "use appropriate representations and models". These applets are used to meet all of these principles with dynamic demonstrations of difficult concepts. In Section 2, we present the applets and give instructions for how they can be used. In Section 3, ideas are provided about how to implement the applets into an introductory class, and Section 4 provides some suggested exercises that may help to further solidify the concept for students through individual exploration and manipulation. Finally in Section 5, we provide some additional uses for the applets that we have found in a consulting environment.

2. The Hypothesis Testing and Power Applets

The Hypothesis Testing and Power Applets are part of a series of Java Applets called Statistical Java, (http://www.stat.vt.edu/~sundar/java/applets) which are being developed and maintained at Virginia Tech. There are a total of seven applets that will be presented in this paper, divided into two sets. The first set with 4 applets is labeled "Hypothesis Testing" and deals with the simplest case of a test: a one-sided hypothesis test of the mean for a one-sample of data from a Normal distribution when the variance is known. There are three parameters which affect this test - the sample size, the proposed significance level (alpha) and the difference between the hypothesized value of the mean and the "true" value. The first three of the applets in this set each allow variation of one of the parameters, while the fourth allows simultaneous variation of all three parameters. The second set of 3 applets is labeled "Power" and deals with the more general cases, commonly dealt with in many introductory courses. The first applet deals with z-tests on one or two means; the second with t-tests on one or two means; and the third with tests on one or two proportions.

Each of the sets of applets can be found under Statistical Theory on the webpage listed above and include a "Main Page" with an overview of the statistical concepts presented and "Applet Instructions" about the detailed functioning of the applets.

The "Hypothesis Testing" group of applets presents similar content to the applets in Aberson et al. (2002) with a restricted case: a one-sample test of the mean for the null hypothesis $H_0: \mu = 10$ versus the one-sided alternative $H_A: \mu > 10$ when the underlying population is approximately Normal and the population variance is known to be $s^2 = 1$. Suggestions of how to make the example relevant to student experience are given in the subsequent section.

Since students frequently are overwhelmed with the interrelationship between the factors that can affect power, separate applets show the effect of changing the sample size, the alpha-level and the "true" value of μ . Figure 1 shows a screen shot of the applet for changing delta, Δ , the difference between the mean value assumed under the null hypothesis and the actual value if the alternative hypothesis is correct. The other parameters affecting power are held fixed (sample size of n = 5, and proposed significance level of $\alpha = 0.05$). The cutoff value for the observed sample mean, $\overline{\chi}_{\alpha}$, is shown, in order to demonstrate that if a sample mean of at least 10.736 is observed it will lead to a decision to reject \overline{H}_0 . The slide bar allows students to change this difference and see the effect on the Type I Error rate, Type II Error rate and Power, both numerically on the right hand panel and graphically in terms of changing areas in the main figure. The user can vary the value of delta between -1 and 5 using the slidebar. Separate applets exist for seeing the effect of changing sample size (from 1 to 30 with the other parameters fixed at $\Delta = 1$ and $\alpha = 0.05$) and for changing alpha (from 0.005 to 0.25 with $\Delta = 1$ and n = 5).

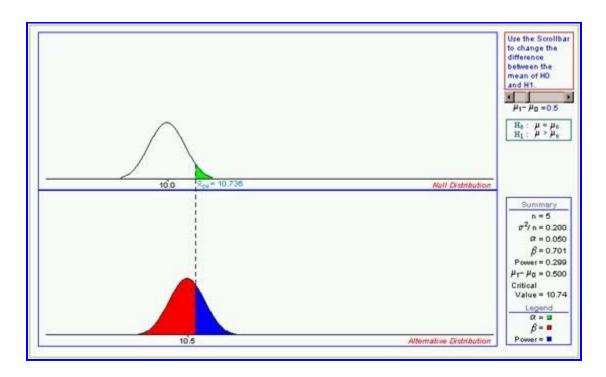


Figure 1. Applet to demonstrate the effect of changing Δ .

Finally, the fourth applet in the set on "Hypothesis Testing" allows students to simultaneously manipulate all three of \triangle (from -1 to 5), α (from 0.005 to 0.25) and n (from 1 to 30) to see their interdependence. This applet is shown in Figure 2. For a demonstration of the actual applet use the following link, (http://ww2.amstat.org/publications/jse/v11n3/java/Hypothesis/). These applets are similar in content to those developed by Aberson et al. (2002) and West and Odgen (1998) but present the concepts in a slightly different manner with some dynamic graphical plots. In particular, the ability to restrict students to experiment with a single parameter is helpful to avoid information overload. It also allows instructors greater ability to emphasize the impact of changing each parameter individually before combining the results.

Once students have been exposed to the basic ideas of hypothesis testing, the restrictions of the one-sample one-sided hypothesis test of a particular value are dropped and the Power applets allow students to explore a much broader range of applications, including one-sided and two-sided hypothesis tests, one- and two-sample problems, as well as testing either the mean or proportions. These applets cover a larger breadth of situations than Aberson et al. (2002) and with different graphical summaries of the concepts involved. As well, the applets present a smaller number of experimental situations than the site by Russ Lenth (http://www.stat.uiowa.edu/~rlenth/Power), but with greater emphasis on dynamic graphical summaries of the concepts.

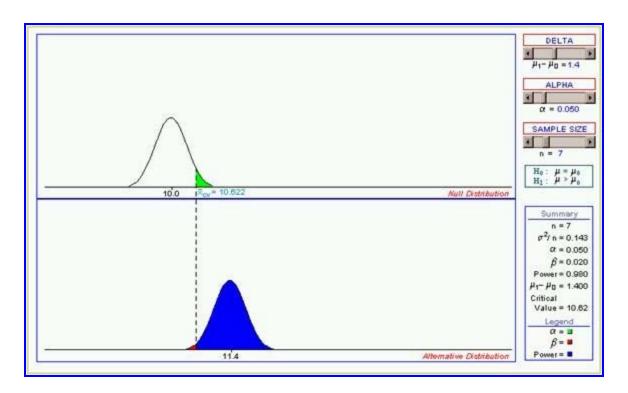


Figure 2. Applet to demonstrate the effect of changing Δ , α , and n.

The set of applets on "Power" contains three applets and covers tests of proportions, z-tests and t-tests, respectively. Within each applet there is flexibility to select the number of samples as well as which alternative hypothesis (less than, greater than or non equal) will be tested. In each of these applets, the distributions under the null and alternative hypotheses shown as in the "Hypothesis Testing" applets, but also a curve is given above that shows the changing level of Power under different alternative hypothesis conditions. Figure 3 shows a screen shot of the t-test applet for the one-sample case testing the two-sided alternative with an observed user-selected sample standard deviation of s = 8, sample size n = 16 and proposed significance level $\alpha = 0.01$. The actual applet is located at http://ww2.amstat.org/publications/jse/v11n3/java/Power.

The scale on the x-axis changes for different scenarios to show the complete range of power values possible. Selecting the type of test (one- or two-sample as well as the type of alternative hypothesis) is done with a click on the appropriate circle, and parameter values for s, n and α is made by clicking in the appropriate box and typing in the desired value. The lower graph remains to show comparisons between the null and alternative hypotheses for any selection of parameter values. In addition, the upper graph now provides curves for the power levels for different Δ values. Numerical values for α given Δ and its associated power are available by moving the cursor (and crosshairs) to the desired location for different alternative hypothesis parameter values. The lower plot dynamically moves to reflect the changing values of the "true" mean under H_A .

Two additional curves in the upper plot, one on each side of the primary power curve, show the effect of changing the sample size by 4 observations. Hence in <u>Figure 3</u> we can see the power curves for samples of size 12, 16 (in darker line) and 20. This enables the students to easily see the effect of changing sample size on power in a single graph. Furthermore, single-clicking on the applet toggles the movement of the crosshairs on a given delta, while double-clicking on the applet will move the crosshairs vertically from one power curve to the next.

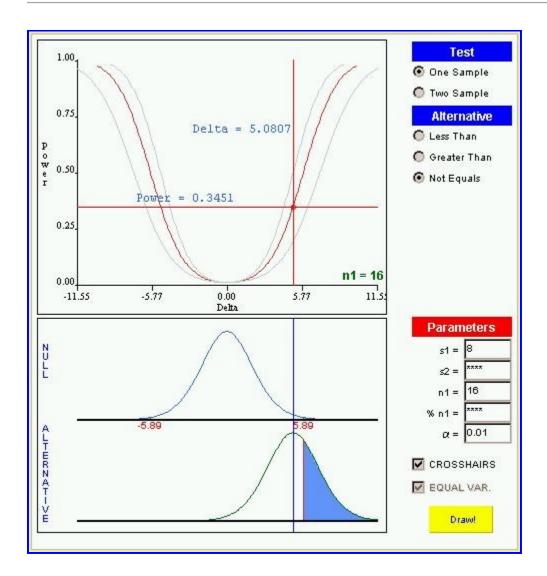


Figure 3. Applet to demonstrate and quantify the power level for a one-sample *t*-test.

For the z-test, which assumes that the standard deviations for the one- or two-samples are known, the parameters required to draw the graphs are σ_1 , σ_2 , n_1 , % n_1 , (which is the percentage of sample size 1 desired for the second sample with default set at 100%) and α . For the one-sample case, σ_2 , and % n_1 are not required by the user and hence are not available.

For the *t*-test, which assumes that the standard deviations for the one- or two-samples are unknown, the parameters required to draw the graphs are s_1 , s_2 (the observed sample standard deviations), n_1 , $%n_1$, and α . Again for the one-sample case, only s_1 , n_1 , and α are required. In addition, for the two-sample case there is an option to assume equal standard deviations for the two-samples (which is the default, as it is the standard approach for most introductory courses) or to work with non-equal sample standard deviations using the more conservative Satterthwaite approximation.

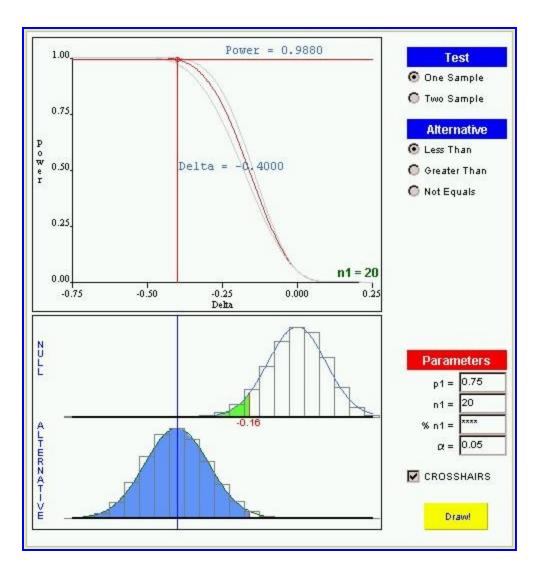


Figure 4. Applet for power of two-sample proportions test.

For the proportions applets, there are a number of differences. First, on the bottom plots, a histogram is shown for the true probabilities for each number of successes based on the actual binomial distribution, in addition to the superimposed Normal distribution that is used for the calculations. Secondly, there are only three and four parameters that are required, respectively, for the one- and two-sample tests: p_1 , n_1 , % n_1 and α . This is, of course, because the variance of the binomial (and also the matching approximated Normal) is completely specified by the proportion parameter. More advanced students will notice how the shape of the distribution associated with the alternative hypothesis changes for different selected values. Also, the range of the power curves is restricted to match the possible values of $0 \le p \le 1$. Finally, for the two-sample case, p_1 is selected by the user and as Δ is changed, p_2 is altered which affects the variances of both the distributions for H_0 and H_A . Figure 4 shows a screen shot of the applet for the proportions, while the actual applet can be accessed at http://ww2.amstat.org/publications/jse/v11n3/java/Power. Notice here that since p_1 is set at 0.75, the delta value ranges from -0.75 (which corresponds to $p_1 = 0$) to 0.25 ($p_1 = 1$).

3. Implementing the Applets into an Introductory Course

In this section, some ideas are given as to how the applets discussed in the previous section might be incorporated into a standard introductory statistics course. For some courses with only a short amount of time available to be devoted to this topic, presenting only the set of "Hypothesis Testing" applets will still allow students the ability to better visualize the relationships between sample size, alpha-level, delta value and the observed power. This could be done in about 30 minutes, with the instructor demonstrating the applets in class. However, we have incorporated both sets of applets into lectures in several of our courses in various ways using a total of 50-75 minutes. In addition, the students are asked to use the applets on their own with some assignment problems. Having the students work with the applets themselves has a dramatic impact on helping them to reinforce the concepts and to better prepare them to be able to solve problems that they are likely to encounter after completing the introductory statistics course

When we introduce the applets into a course, we have already developed hypothesis testing and taught the students how to perform the most straightforward of the tests, the one-sample z-test with variance assumed known. The concepts of Type I and II error have been presented to begin the process of understanding the trade-offs in the hypothesis testing decision-making process. At this point, a particular example appropriate to the background of the students in the course is presented which tests the null hypothesis $H_0: \mu = 10$ versus the alternative $H_A: \mu > 10$ when the population variance is known to be $\sigma^2 = 1$. For example, for a course dominated by engineering students, the example may involve testing a new glue's average strength. For a course with ecology majors, the test may study average fish size in a given river; for biology students, the test may relate to the average drug response. In this way, the students can potentially identify with the example as considerable time will be spent examining it and discussing the repercussion of the types of errors.

We first present the hypothesis testing applet which allows \triangle to be changed, and discuss what would represent making a Type I error or a Type II error in terms of different possible \triangle values. The practical implications of both types of errors are discussed in terms of the particular example. In addition, attention is given to the interpretation of what the power quantity means for the given problem, where only a single experiment will be performed. For this applet, it should be emphasized to the students that the true value of \triangle is typically unknown and is not user specified or controlled. In this sense, the applet shows a variety of scenarios, where only one is correct, but which one is unknown.

Next the applet for changing α is presented, which allows for a quantification of the trade-offs between Type I and Type II error, under the particular scenario. Key points here are that this is a user-specified and controlled parameter, but there are a limited number of commonly accepted choices. In addition, students quickly realize that changes to this parameter frequently do not make dramatic differences to the outcome of the test.

Next, the applet for changing the sample size, n, is presented. Here students see the effect of changing sample size on the distribution under the null and alternative hypotheses, and consequently on power. This is the major tool available to experiment designers to influence power. We have found that this presentation can effectively be done in about 30 minutes of lecture time. At the conclusion of these demonstrations by the instructor, students are given an opportunity to experiment with the applets on their own in the context of simple assignment questions that reinforce the effect of changing Δ , α and n, both individually and all three simultaneously (with the fourth applet).

We then continue the regular lecture sequence to teach hypothesis testing for the variance unknown case (t-tests), testing of proportions and the two-sample cases. At the conclusion of the hypothesis testing

unit, we return to the set of "Power" applets. A brief review of the lower portion of the plots is generally sufficient to refresh the students' memory on the basics power. We then change the presentation of the material to focus on designing an experiment to answer a particular question. We begin with a discussion of the difference between practical importance and statistical significance (what is an important scientific result versus what difference is unlikely to have occurred by chance). We present this in the context of a particular problem, with specified possible outcomes if we reject the null hypothesis - we find a statistically significant difference that is important, we find a statistically significant result that is unimportant. Students are asked to anticipate how each of these might occur, and what the associated costs are of each.

The set of "Power" applets is then presented, with some explanation of what the new upper curve summarizes. Then a number of examples are presented to show how the applets can be used to answer questions for particular problems. Details of the types of questions possible are given in the following section. After an in-class demonstration, students are encouraged to experiment with the applets themselves by completing a comprehensive assignment.

The power applets can also be related to other resources available on the Statistical Java website. Specifically, the Power applet for the one sample test on proportions can be connected to both the Central Limit Theorem applet and the Confidence Intervals applets. Because the tests on proportions use a Normal approximation and rely on large sample sizes it may be helpful to a student to see the effects of smaller sample sizes on any inference. In particular, The Central Limit Theorem applet allows the user to visualize when the distribution of the sample proportion approaches a Normal distribution by increasing the sample size. The Confidence Interval applets demonstrate the effects of changing the sample size on confidence interval width for a single proportion. Connecting these applets will reinforce the ideas of power, hypothesis testing and confidence intervals.

Other applets on the STATISTICAL JAVA website that relate to the Power applets include the Control Chart applets and the Distribution applets. The Control Chart applets demonstrate the Shewhart control chart for a sample mean assuming a Normal distribution and a known process variance. Since control charts help visualize sequential hypothesis tests, the Power applets help demonstrate the effects of using 3-sigma limits on the probability of detecting a signal. More indirectly, the Distribution and Probability applets allow a student to see the shapes of different distributions, including the Binomial, Normal, and T, which are all used in the Power applets.

4. Suggested Exercises

Based on the set of "Hypothesis Testing" applets, the following is a sample of questions to help
reinforce the ideas presented in class. For an experiment (add relevant context here for different classes)
with sample size, and chosen α of to test the hypotheses $H_0: \mu = 10_{\text{VS}} H_A: \mu > 10$ when the
population variance is known to be $\sigma^2 = 1$, answer the following questions:

- a. What is the cutoff value for this test? What does this mean?
- b. What is the probability of making a Type I error, and describe what would need to happen in order for us to make this type of error?
- c. If we obtained a sample mean of _____, what would our conclusion for the test be?
- d. How likely is it that an experimenter will reject H_0 if the true population mean value is ____? What is the probability of making a Type II error?

Additional questions could consider how much the power changes if we change the α -level or sample size.

A wider variety of questions is possible for the set of "Power" applets, both because of the number of different scenarios permissible, as well as the additional quantitative summaries. Problems should be posed in terms of real world situations relevant to the students' experiences and interests and in terms of solving a problem of practical importance. The style of question is appropriate for each of the *z*-, *t*- and proportions tests, and a variety of one- and two-sided hypotheses should be used. For brevity and generality, just the skeleton of the questions is presented. Here are a number of different questions, in roughly increasing order of complexity.

- i. For a given \triangle , α and n, find the power of the test.
- ii. For a given α and n, find the minimal Δ that will lead to you rejecting the null hypothesis ____% of the time.
- iii. For a given \triangle and α , find the sample size that would lead you to reject H_0 at least _____% of the time. (This requires the students to be more interactive than either of the two previous questions as they need to experiment with different sample sizes.)
- iv. For a given \triangle and n, how much gain in power is obtained by changing from $\alpha = \underline{\hspace{0.5cm}}$ to $\underline{\hspace{0.5cm}}$? Which α do you feel is more appropriate, based on your answer?
- v. For a paired *t*-test, with observed data given and $\alpha = ___$, find the power of the test if the true difference between means is ____. For this particular set of data, was the correct decision made as a result of the hypothesis test? (This requires considerable consolidation of material as students need to use the connection between the paired *t*-test and the one-sample test.)
- vi. What sample size would be required for an experiment (with given Δ and α) to have power at least as good as a scenario with Δ , α and n given? (This allows students to find a power first, and then use this power to find a sample size.)

By mixing and matching the style of question above with different types of tests and hypothesis, a rich balance of problems can be obtained that are computationally very light, but conceptually very challenging. One key aspect of these questions should be to pose questions in the context of real world applications, that will force students to make decisions about which of the tests is appropriate for a given scenario.

5. Conclusions and Comments

We have found that students in introductory statistics classes react very positively to the applets, both in terms of enjoying being able to experiment with them as well as being better able to discuss the concepts relating to power. Anecdotally, students' performance on test questions related to the concepts of power, sample size and hypothesis testing in recent years has improved. On end-of-term evaluations, students frequently cite the use of the applets as one of their favorite parts of the course and comment that they found them very helpful. Because of ethical guidelines at our institution regarding only a fraction of students enrolled in a course having access to teaching materials, and the analysis implications of student self-selection, no formal testing of the effectiveness of the applets has been attempted. The applets can be used in a variety of ways and have been deliberately designed to allow for maximum flexibility.

In addition to using the set of "Power" applets in first statistics courses, we have also used them as a refresher in our advanced undergraduate and graduate statistical consulting class. Many of the statisticians in these classes are comfortable with the concept of hypothesis testing, but do not have much experience with applying the ideas to designing experiments. The applets provide good focus for discussing how to guide a consulting client through experimental design considerations to obtain a sensible sample size for the desired practical implications of the experiment. This connection to reality is helpful for statistically-savvy but experimentally-naïve students.

The "Power" applets have also been used extensively with individual consulting clients to show them the impact of different sample sizes for experiments that they are planning to run. By removing the computational burden, it is easy to discuss power with clients in a real-time environment. Several experiments have been substantially redesigned, some with an increase in sample size, and some abandoned completely, once the power for the experiment is quantified.

We believe that the applets will be an easily accessible tool for instructors of introductory statistics courses to incorporate into their lectures and assignments to help students gain a better working understanding of the concepts and implications of power and sample size for data collection.

References

Aberson, C. L., Berger, D. E., Healy, M. R., and Romero, V. L. (2002) "An Interactive Tutorial for Teaching Statistical Power," *Journal of Statistics Education* [Online], 10(3). (ww2.amstat.org/publications/jse/v10n3/aberson.html)

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), New York: Academic Press.

Gatti, G. G., and Harwell, M. (1998), "Advantages of Computer Programs over Power Charts for the Estimation of Power," *Journal of Statistics Education* [Online], 6(3). (ww2.amstat.org/publications/jse/v6n3/gatti.html)

Nickerson, R. S. (1995), "Can Technology Help Teach for Understanding" in *Software Goes to School*, eds. D. N. Perkins, J. L. Schwartz, M. M. West, and M. S. Wiske, Oxford: Oxford University Press.

West, R. W., and Ogden, R. T. (1998), "Interactive Demonstrations for Statistics Education on the World Wide Web," *Journal of Statistics Education* [Online], 6(3). (ww2.amstat.org/publications/jse/v6n3/west.html)

C. M. Anderson-Cook Department of Statistics Virginia Tech Blacksburg, VA 24061 USA candcook@vt.edu

Sundar Dorai-Raj

PDF Solutions, Inc.
Dallas TX 75082
USA
sundar.dorai-raj@pdf.com

Volume 11 (2003) | Archive | Index | Data Archive | Information Service | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Home Page | Contact JSE | ASA Publications