Java Applets for Sample Size

Russell V. Lenth
Department of Statistics and Actuarial Science
University of Iowa
Iowa City, IA 52242
russell-lenth@uiowa.edu
Joint Statistical Meetings, Atlanta, August, 2001

This handout briefly describes some Java applets for power and sample size. Additional comments are made regarding concepts, approaches, and practices of sample-size determination. More information is available from the author's web site.

Software features

The power-analysis software runs on almost any system. The easiest way to run it is to simply connect to the author's web site, `http://www.stat.uiowa.edu/~rlenth/Power/`, using a Java-capable web browser. It can also be run without an Internet connection by obtaining the file `piface.jar` (downloadable from the same web site) and having the Java Runtime Environment (downloadable from `www.sun.com`) installed on your system. All of this software is free.

Several power-analysis scenarios are available by selection from a list. Each brings up a dialog box on the screen. By design, these dialogs are highly interactive. Most of them feature sliders that can be manipulated and the results of those manipulations are immediately seen. Sliders are convertible to text-entry fields when that is desired. Every dialog has available a linked graphics dialog; if one or more graphs are showing, then they too change dynamically as sliders and other controls are manipulated. Graphs are designed for interaction rather than presentation, but there is a provision for retrieving the values plotted and pasting them into another application.

Significance and equivalence—together

One of the most common types of questions I get concerning use of these applets concerns using them to compute power retrospectively. This issue comes up when data have already been collected and a hypothesis has been tested, with a "non-significant" result. Then one might ask questions such as, "is the effect really pretty negligible?" and will try to address it by doing a power calculation.

This use of retrospective power is problematic from an inferential standpoint; see Hoenig and Heisey (2001), Lenth (2001) for explanations, but the main issue is that you're trying to reverse the null and alternative hypotheses while still using the same statistical test and critical-region boundaries. When the hypotheses are reversed, we have a test of *equivalence*, and there is a corresponding formal approach. For example, in the two-sample t-testing situation, here are formulations of tests of significance and equivalence:

Significance	**Equivalence**
$H_{0S} : \lvert\mu_1 - \mu_2\rvert = 0$	$H_{0E} : \lvert\mu_1 - \mu_2\rvert \geq \tau$
$H_{1S} : \lvert\mu_1 - \mu_2\rvert > 0$	$H_{1E} : \lvert\mu_1 - \mu_2\rvert < \tau$

where τ is some specified threshold. Schuirmann (1987) shows that a (slightly) conservative α-level test of H_{0E} versus H_{1E} is obtained by rejecting H_{0E} if and only if a $1-2\alpha$ confidence interval for $\mu_1 - \mu_2$ lies entirely in the interval $[-\tau, +\tau]$. This is really a combination of two one-tailed t tests, each at level α; no adjustment is needed for multiple testing, because the two null hypotheses are disjoint.

Now, here's an interesting point: H_{0S} and H_{0E} are also disjoint. So by what was just said, we can test both significance and equivalence, each at level α, and the overall type-I error probability for the two tests is still protected at α. This puts you on very firm inferential ground for establishing both positive and negative findings, and avoids the pitfalls of retrospective power.

The time for power calculations is before the study is conducted. At this point you can plan the study so that there is adequate sample size for testing *both* significance and equivalence. My Java applets make it quite easy to do that in the two-sample t setting, and plans are underway to provide similar capabilities for other statistical tests.

The ROC approach

One traditional approach to sample-size determination problem entails specifying a target effect size, values of one or more related parameters (e.g., variances), a significance level α, and the desired power of detecting the stated effect size. Then the sample size is manipulated so that the desired power is achieved. Most of the Java dialogs are designed with this underlying paradigm.

We can simplify this procedure a little bit by borrowing an idea from medical diagnostics. A plot of sensitivity versus $1-$ specificity of a diagnostic test is called an *ROC curve*, and the area under the curve (*AUC*) is considered an overall measure of its diagnostic capability; see, for example, Hanley and McNeil (1982). In the context of hypothesis testing, a plot of power versus α is an ROC curve, and *AUC* is just the average power over all α. More important, perhaps, is the following interpretation: Pretend that the study can be conducted independently in two parallel universes, one in which the null hypothesis is true and the other in which a non-null effect of stated size holds. Let T_0 and T_1 be the respective test statistics from these hypothetical studies. Then it can be shown that $AUC = \Pr(T_1 \text{ is more significant than } T_0)$. Again, this is comparable to the medical-diagnostic interpretation of *AUC*.

By considering *AUC* as the criterion in a sample-size problem, there is one fewer parameter to specify. My two-sample *t*-test dialog provides for this option—in both the significance and equivalence tests.

Mixed-effects ANOVA

The most complex of the applets is the one for mixed-effect analysis of variance. Any balanced mixed-effects design (with independent terms) can be studied. You may either select a design from a list, or specify the model in a text format (similar to that of SAS, but with terms delineated by $+$ signs, which makes it easier to parse). One may focus either on ANOVA F tests or (as I recommend for fixed terms) on tests of comparisons or contrasts. The dialogs for both approaches can be used simultaneously, and common parameters are linked together.

In the F-test dialog, effect sizes are specified using standard deviations—because these make sense for both fixed and random effects. The "unrestricted model" is used, and where necessary, pseudo-F tests are constructed using the Satterthwaite method. To help specify meaningful SDs for fixed effects, auxiliary dialogs are available from the Options menu that display linked, manipulable dotplots (for main effects) or interaction plots (for two-way interactions).

In the comparisons/contrasts dialog, the user may choose from among a variety of adjustments for multiple testing. Power is computed on a per-contrast basis. In certain designs (e.g. split-plots), the variance of a contrast of cell means is different depending on whether or not it is restricted to the same levels of a blocking factor; this possibility is supported.

With both approaches, useful reports are available from the Options menu, including a summary of computed powers, expected mean squares, and coefficients of constructed variance estimators.

Developing your own applets

(This section is for Java programmers only!) If you wish, it is possible to write your own Java code that makes use of the machinery behind these applets. Simply download `piface.jar` and put it somewhere in your `CLASSPATH`. (Alternatively, just specify `ARCHIVE=http://www.stat.uiowa.edu/~rlenth/Power/piface.jar` in your HTML `APPLET` specification.)

A user interface similar to those in these sample-size applets is very easily obtainable by extending the class `rvl.piface.Piface`. All you need to do is write a `gui()` method that sets up the user interface (often, one line of code per component), and a `click()` method that is called whenever a user action takes place. `piface.jar` also contains a number of classes with static methods for cdfs, quantiles, and power functions of standard distributions, and extras such as root-finding and numerical integration. Documentation is available from the author's web site.

Useful references

Hanley, J. A. and McNeil, B. J. (1982), "The meaning and use of the area under a Receiver operating characteristic (ROC) curve," *Radiology*, 143, 29–36.

Hoenig, J. M. and Heisey, D. M. (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations in Data Analysis," *The American Statistician*, 55, 19–24.

Horstmann, C. S. and Cornell, G. (1997), *Core Java 1.1*, Sun Microsystems Press, Mountain View, CA, Two-volume set.

Lenth, R. (2001), "Some Practical Guidelines for Effective Sample-Size Determination," *The American Statistician*, 55, 187–193.

Odeh, R. E. and Fox, M. (1991), *Sample Size Choice: Charts for Experiments with Linear Models*, Marcel Dekker, New York, second edn.

Satterthwaite, F. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110–114.

Schuirmann, D. (1987), "A compromise test for equivalence of average bioavailability," *ASA Proceedings of the Biopharmaceutical Section*, 1987, 137–142.

Searle, S. R. (1971), *Linear Models*, Wiley, New York.

Thomas, L. (1997), "Retrospective Power Analysis," *Conservation Biology*, 11, 276–280.

Zumbo, B. D. and Hubley, A. M. (1998), "A note on misconceptions concerning prospective and retrospective power," *Journal of the Royal Statistical Society, Series D*, 47, 385–388.