Statistical power calculations[1]

R. V. Lenth[2]

Department of Statistics and Actuarial Science, University of Iowa, Iowa City 52242

ABSTRACT: This article focuses on how to do meaningful power calculations and sample-size determination for common study designs. There are 3 important guiding principles. First, certain types of retrospective power calculations should be avoided, because they add no new information to an analysis. Second, effect size should be specified on the actual scale of measurement, not on a standardized scale. Third, rarely can a definitive study be done without first doing a pilot study. Some simple examples as well as a complex example are given. Power calculations are illustrated using Java applets developed by the author.

Key words: sample size, statistical power

J. Anim. Sci. 2007. 85(E. Suppl.):E24–E29
doi:10.2527/jas.2006-449

INTRODUCTION

Obtaining an appropriate sample size is an important aspect of planning a statistical study. This article outlines some important concepts and techniques for this purpose. Emphasis is on the power approach, which is described in the next section. Some simple examples are provided in the subsequent sections “One Proportion” and “Two-Sample *t*-Test.” The section “What Power Isn’t” discusses some common misconceptions about power and effect size, and the section “Practical Recommendations” suggests some practical issues and the role and need for a pilot study. Finally, I give an example of a relatively complex experimental plan, in which there are multiple study goals and the pilot study has a different experimental design than the planned study.

POWER

Power is the probability of obtaining a statistically significant result using a statistical test. It depends on the significance level (α) of the test, the sample size (n), the actual effect size (θ; on the original scale of measurement), and possibly other “nuisance” parameters such as the error SD, σ. Power is useful in several ways for planning a future study. For example, it is useful for deciding the sample size for a given effect size of clinical importance or for evaluating the power of a planned study when the sample size has been dictated by budget constraints.

Actually calculating the power of a test often requires the use of noncentral distributions such as the noncentral *t*. These are complicated calculations, and, hence, very few closed-form exact sample-size formulas exist. Although there is an abundance of approximate formulas, they are less and less necessary due to the availability of computer software that can do the exact noncentral calculations. Commercial standalone software includes nQuery Advisor (http://www.statsol.ie/nquery/nquery.htm) and PASS (http://www.ncss.com). Many general-purpose statistical programs (e.g., SAS, Minitab, and others) include sample-size procedures. There also is a variety of freeware for sample-size computation, such as Piface (Lenth, 2006). The latter are Java applets on my own Web site, and they are used for the illustrations in this article. When using software products for sample size, one must be cautious, because some programs still incorporate old (and sometimes poor) approximations. The software mentioned above all perform exact calculations.

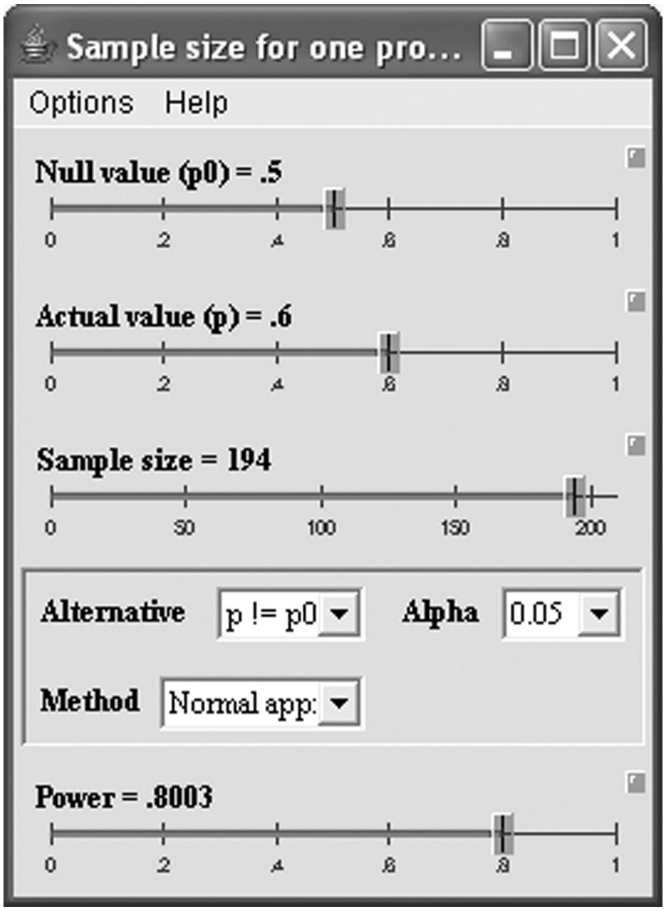
ONE PROPORTION

One of the simplest statistical studies involves testing a proportion. For example, one may plan to collect data for a paired comparison of a treatment and a control condition and use the sign test to evaluate its significance. To do the sign test, we simply go through each pair and record a “+” if the treatment is greater than the control, and a “–” otherwise. Under the null

[1] Presented at the ADSA-ASAS Joint Annual Meeting, Triennial Reproduction Symposium: Molecular Techniques and Statistics, Minneapolis, MN, July 2006.

[2] Corresponding author: russell-lenth@uiowa.edu

Received July 7, 2006.
Accepted October 13, 2006



Null value (p0) = .5

Actual value (p) = .6

Sample size = 194

Alternative p != p0

Alpha 0.05

Method Normal app:

Power = .8003

Figure 1. Piface implementation of the sign-test example.

hypothesis that there is no difference between treatment and control, we would expect one-half plusses and one-half minuses. Accordingly, let p denote the probability that treatment is greater than control. The 2-sided sign test can then be formulated as a test of the null hypothesis, H_0: $p = ½$, against the alternative, H_1: $p \neq ½$. If $\hat{p}$ denotes the observed proportion of plusses based on n treatment-control pairs, then the normal approximation yields the test statistic $z = (p - 0.5)/\sqrt{0.5 \times 0.5/n}$.

For a significance level of $\alpha = 0.05$, we deem $\hat{p}$ significantly different from one-half (and, hence, the treatment differs significantly from the control) if $|z| > z_{\alpha/2}$, the critical value that cuts off a tail of area $\alpha/2$ on the standard normal distribution.

To plan an experiment that uses this test, we need to specify what value of p would be regarded as clinically significant; that is, how far must p deviate from one-half in order to say that the treatment differs from the control in a meaningful way? Suppose that we discuss this matter with the research team, and they decide that $p = 0.6$ or more or $p = 0.4$ or less constitutes a clinically important difference from 0.5. The sample-size problem is then to find the value of n such that the power of the test is reasonably high (say 0.80 or 80%) when $p = 0.6$. This is a simple matter using Piface, as illustrated in Figure 1. Sliders are used to set the null value of $p = 0.5$, the alternative of interest at $p = 0.6$, and other graphical elements to set the significance level, the 2-sided alternative, and to use the normal approximation. Then selecting a power of 0.8 on the bottom slider yields $n \approx 200$.

TWO-SAMPLE *t*-TEST

In this section, I discuss planning a study involving a 2-sample t-test to compare 2 means. We want to compare 2 treatment conditions where data are collected in independent samples, and it is decided that a 15% difference (factor of 1.15) is clinically important. Often, when a relative or percentage difference is considered, it is appropriate to analyze the data on the logarithmic scale. Moreover, a percentage difference of the original values translates to a shift difference on the logarithmic scale. Note that

$$\ln 1.15 \approx 0.14$$

and

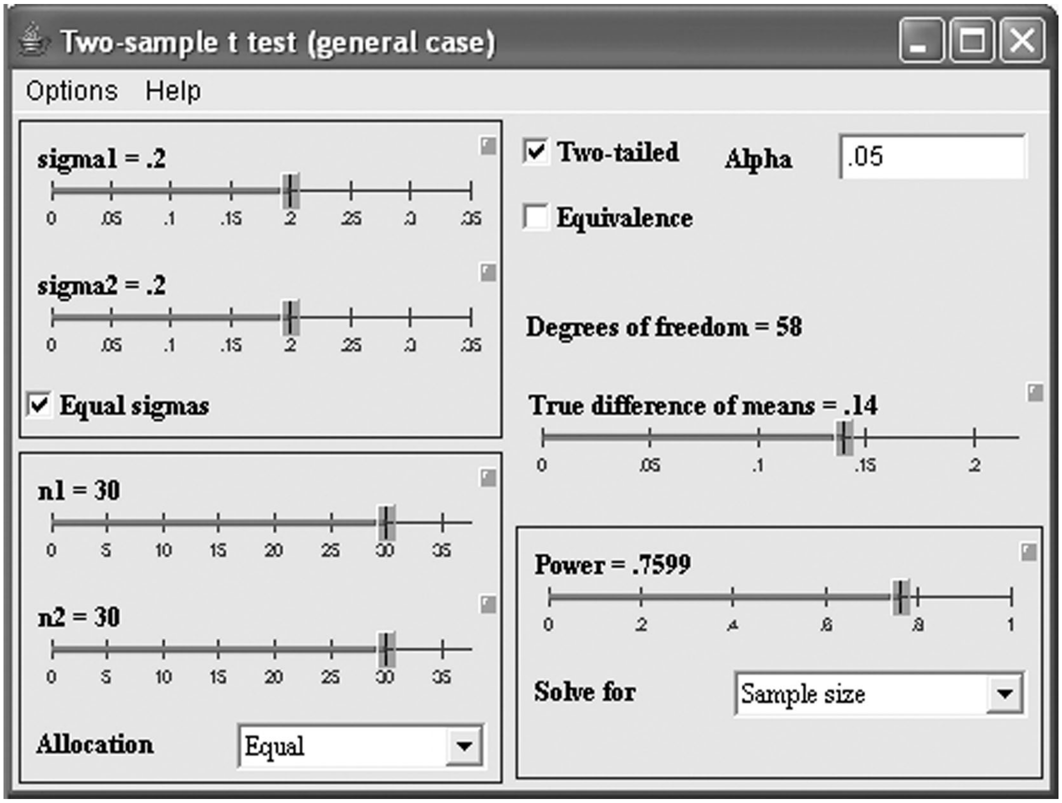
$$\ln 1/1.15 \approx -0.14.$$

Thus, a 15% difference translates to a difference of ±0.14 on the natural-logarithm scale.

Unlike the sign-test scenario, we also need an estimate of the error SD, σ. Suppose that pilot data (analyzed on the ln scale) suggest that $\sigma \approx 0.20$. Suppose also that the budget is just sufficient to collect 30 observations per condition. Figure 2 shows the Piface dialog for this experiment. On the left-hand side, the values $\sigma = 0.2$ and $n = 30$ are entered for both conditions. On the right-hand side, we enter $\alpha = 0.05$, that a 2-tailed test is desired, and that the difference of interest is 0.14. We find that the power is about 0.76 and thus that the budgeted sample size is reasonably adequate. If the power had come out small, one could argue for a bigger budget, or discover what difference of means could be detected with $n = 30$ in each group, and reevaluate whether the planned experiment is worth carrying out.

WHAT POWER ISN'T

Before proceeding to a more complex example, it is worth discussing some common mistakes related to power computations. First of all, power is not useful in data analysis; it is useful for planning a future study. One common, but misguided, practice when a result is found nonsignificant is to compute the power retrospectively; that is, the power based on all the observed information (sample size, estimated effect, estimated error SD, etc.). This can be shown to be merely a function of the P-value of the test; when $P < \alpha$, the retrospective



sigma1 = .2

sigma2 = .2

☑ Equal sigmas

n1 = 30

n2 = 30

Allocation: Equal

☑ Two-tailed Alpha .05

□ Equivalence

Degrees of freedom = 58

True difference of means = .14

Power = .7599

Solve for: Sample size

Figure 2. Piface interface for a 2-sample *t*-test.

power will be more than 50%, and when the *P*-value is larger, the retrospective power becomes small. Thus, it adds no new information to an analysis. A good discussion of the reasons not to use retrospective power is provided in Hoenig and Heisey (2001).

I would go beyond what Hoenig and Heisey (2001) say and argue that the traditional way of computing retrospective power is incorrect, because it ignores available information. Given all the information (i.e., sample size, effect size, etc.) that goes into the calculation, one also can deduce the outcome of the statistical test. If we use that information, the fallacy of retrospective power becomes clear. Because power is the probability of a significant result, the retrospective power is equal to 1 if the result is significant, and 0 if it is nonsignificant. It is certainly easier to compute that way!

There are other ways that power calculations can be used legitimately at the end of a study, namely, to decide what is needed for a future, additional study that yields enough data to serve one's goals. That would involve using what we have learned (e.g., the error SD) and an effect size deemed of clinical importance (not the observed effect size) to see how much additional data are needed.

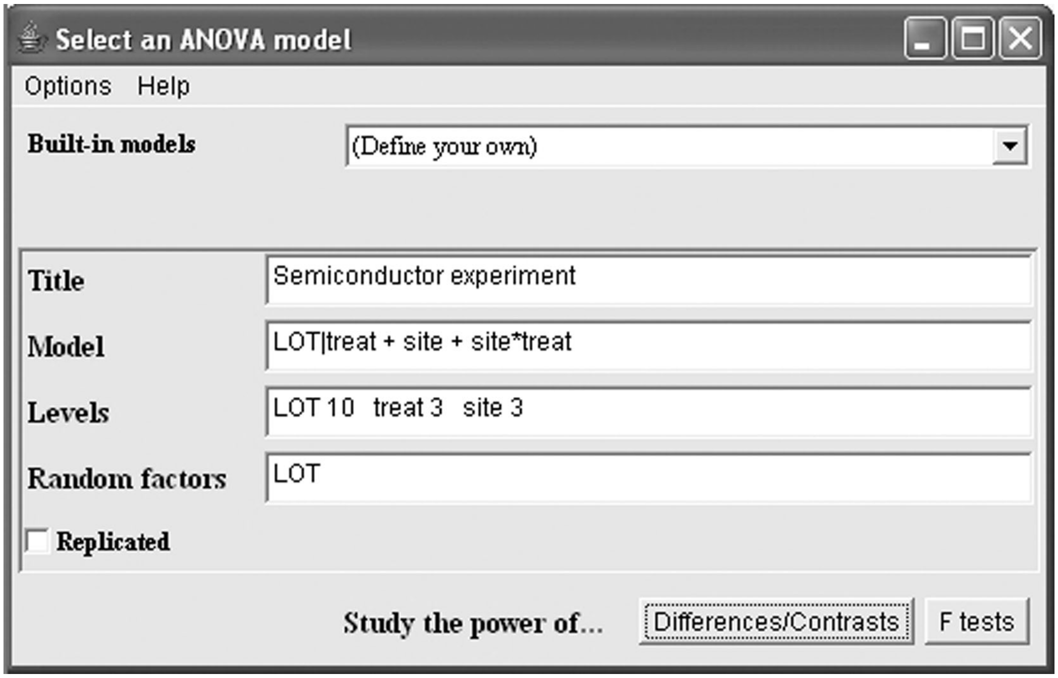
PRACTICAL RECOMMENDATIONS

I believe that the clinical effect size should be specified on the actual measurement scale, not relative to σ, as was done with Cohen's "small," "medium," and "large" effects in popular use (Cohen, 1988). These "T-shirt" effect sizes are based on surveys of the social science literature, and using them simply gives you the same sample size as is commonly used in large, medium, and small published studies in the social sciences. What appears to be a calculation is actually an elaborate way to arrive at a foregone conclusion.

In practice, it can be difficult to have an effective conversation about effect size. Often, a useful line of inquiry for effect size can be approached like this: "How different can these groups be and still be considered practically the same?" As we saw in the *t*-test example, we need an idea of σ for most common analyses. If you truly have no idea of σ, then I would say that you are not ready to do a definitive study and should first do a pilot study to estimate it. Summarizing, there are 2 essential ingredients for power analysis:

1) **Put science before statistics:** This involves a serious discussion of study goals and effects of clinical importance, on the actual scale of measurement.
2) **Pilot study:** For estimating σ and also to check to make sure that the planned procedures actually work.

The discipline of doing power calculations right is ideal preparation for successful grant proposals or convincing management of the worth of your proposed project. For more discussion, see Lenth (2001).



Select an ANOVA model

Built-in models (Define your own)

Title Semiconductor experiment

Model LOT|treat + site + site*treat

Levels LOT 10 treat 3 site 3

Random factors LOT

□ **Replicated**

Study the power of...

Figure 3. Interface to set up the split-plot model.

A SPLIT-PLOT EXPERIMENT

In this section, I illustrate how to approach power calculations in a more complex experiment. The example given is in semiconductor manufacturing. Although this is not an animal science experiment, it is an effective example that is easy to understand, and the same ideas apply to many other scenarios.

The structure of the planned experiment is as follows: n lots of silicon wafers are to be produced. In each lot, 3 wafers will be used as experimental units, 1 for each whole-wafer treatment. We measure oxide thickness, in angstroms, (Å) at each of 3 fixed sites on each wafer. Thus, the experimental design is a split-plot with lots as blocks (a random factor) and 2 fixed factors, treatment (whole-plot or between-wafer) and site (split-plot or within-wafer). The model for the planned experiment will have effects for lot, treatment, lot × treatment, site, treatment × site, and error. Lot, lot × treatment, and error are random effects, whereas treatment, site, and treatment × site are fixed effects.

Table 1. The ANOVA for the semiconductor test described in "A Split-Plot Experiment"

Source	df	Sum of squares	Mean square
Supplier	1	1,830.10	1,830.10
Lot (supplier)	6	7,195.20	1,199.20
Wafer (lot supplier)	16	1,922.67	120.17
Site	2	15.44	7.72
Source × site	2	58.33	29.17
Error	44	529.56	12.04

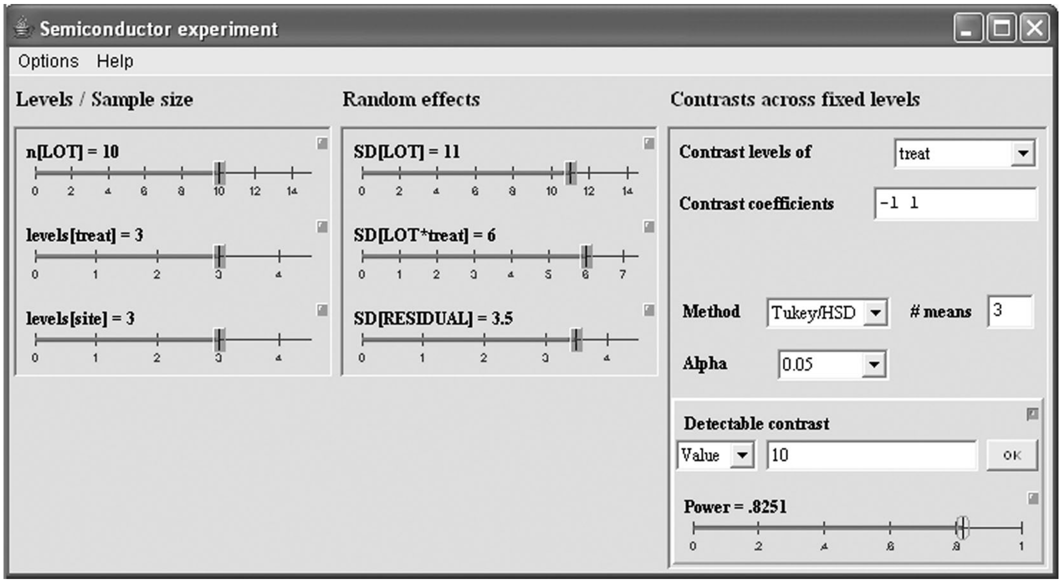
Our design goals are as follows. We want to be able to have at least an 80% power of detecting the following effects, based on tests with significance level 0.05:

- a difference of ±10 Å between 2 treatment means,
- a difference of ±5 Å between 2 site means, and
- a difference of ±15 Å between 2 treatment × site means.

We have available data (Littell et al., 1996) that will help us plan the experiment, but the past experiment has a different design. These previous data comprise 8 lots of 3 wafers each, 4 lots using wafers from 1 supplier and the other 4 using wafers from another supplier; thus, there are 24 wafers altogether. On each wafer, oxide thickness is measured at 3 fixed sites. In this experiment, lots are nested in supplier, and wafers are nested in lot. Site is crossed with the other factors. The ANOVA is shown in Table 1.

Lot, wafer, and error are the 3 random effects. Equating these with their expected mean squares, we obtain the following estimates of their respective variance components: $\hat{\sigma}_L^2 = 119.9$; $\hat{\sigma}_W^2 = 36.04$; and $\hat{\sigma}_E^2 = 12.04$. The estimate $\hat{\sigma}_L \approx 12$ is identified with the lot SD in the planned split-plot experiment; the effect for wafer (lot supplier) corresponds to the lot × treatment effect in the planned experiment, because a combination of lot and treatment identifies a single wafer. Finally, the error SD in the 2 experiments correspond to within-wafer variation. Thus, in spite of the designs being different, there is a 1-to-1 correspondence between the variance components in the past experiment and those in the planned experiment.

Figure 3 shows the Piface dialogue window for specifying an ANOVA model. The model is entered using



n[LOT] = 10
levels[treat] = 3
levels[site] = 3

SD[LOT] = 11
SD[LOT*treat] = 6
SD[RESIDUAL] = 3.5

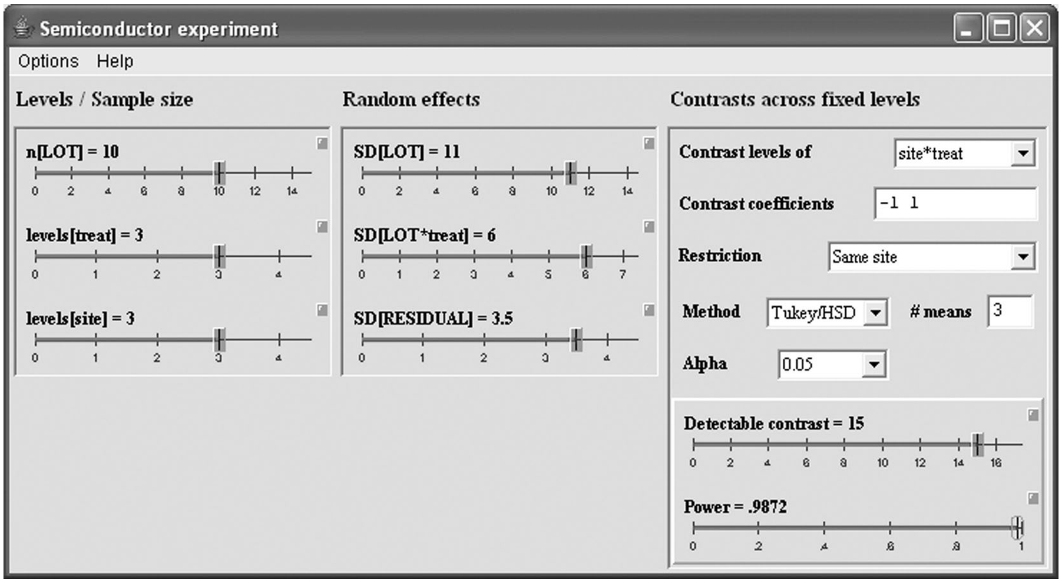
Contrast levels of: treat
Contrast coefficients: -1 1
Method: Tukey/HSD
# means: 3
Alpha: 0.05
Detectable contrast: Value 10
Power = .8251

Figure 4. Interface for treatment comparisons.

SAS-like notation, except the terms are delimited by + signs. The beginning number of levels for each factor and which factors are random also is specified. One has the option of studying the powers of the ANOVA *F*-tests, or *t*-tests of differences, or contrasts among factor levels; we elect the latter.

Figure 4 shows the resulting interface for studying the treatment difference. Factor levels are varied on the left-hand panel, and variance components (actually SD components) are entered on the middle panel; this is where we specify the estimates $\sigma_L = 11$; $\sigma_{LT} = 6$; and $\sigma_E = 3.5$. On the right-hand panel, we specify that we want to compare levels of the treatment factor, enter contrast coefficients of –1 and 1 (this specifies a difference of 2 treatments), that we wish to consider the power of Tukey's honestly significant difference method, and that the difference of interest is 10 Å. When there are 10 lots, the power of this test is about 0.82.

By selecting the site × treatment interaction, we get the dialogue window in Figure 5. The SE of these comparisons differ depending on whether they are between wafers or are within wafers; hence, the additional drop-down list for restrictions on the comparison. We see that with the same number of lots (i.e., 10), for the power of the comparison of 2 treatments on the same site (a between-wafer comparison), the power is very high for the stated difference of interest of 15 Å. The



n[LOT] = 10
levels[treat] = 3
levels[site] = 3

SD[LOT] = 11
SD[LOT*treat] = 6
SD[RESIDUAL] = 3.5

Contrast levels of: site*treat
Contrast coefficients: -1 1
Restriction: Same site
Method: Tukey/HSD
# means: 3
Alpha: 0.05
Detectable contrast = 15
Power = .9872

Figure 5. Interface for treatment × site comparisons.

power of a within-wafer comparison (2 sites with the same treatment) is even greater.

Not shown are the results for the comparisons of sites, but the power is sufficiently high when the true difference is 5 Å, as per study goals. Hence, all of the tests of interest have a power of at least 0.80 when we use 10 lots.

SUMMARY AND CONCLUSIONS

Power calculations are for planning a statistical study, not for analyzing existing data. Although these calculations can be technically messy, this paper shows that with the use of newer interactive software, we can devote most of our attention to the scientific issues to be addressed by a statistical study and perform "what-if" style calculations of power to arrive at a reasonable sample size. This does not mean it is an easy process, just that it is technically fairly simple. The big challenge is establishing communication between scientist(s) and statistician(s) so that the goals of the study can be adequately defined. In addition, a pilot study will often be required before a definitive one can be planned.

LITERATURE CITED

Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Acad. Press, New York, NY.

Hoenig, J. M., and D. M. Heisey. 2001. The abuse of power: The pervasive fallacy of power calculations in data analysis. Am. Stat. 55:19–24.

Lenth, R. V. 2001. Some practical guidelines for effective sample-size determination. Am. Stat. 55:187–193.

Lenth, R. V. 2006. Java applets for power and sample size. http://www.stat.uiowa.edu/~rlenth/Power/. Accessed Oct. 2006.

Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. 1996. Pages 563–564 in SAS System for Mixed Models. SAS Inst. Inc, Cary, NC.