Some Practical Guidelines for Effective Sample-Size Determination

Russell V. Lenth[*]
Department of Statistics
University of Iowa

March 1, 2001

Abstract

Sample-size determination is often an important step in planning a statistical study—and it is usually a difficult one. Among the important hurdles to be surpassed, one must obtain an estimate of one or more error variances, and specify an effect size of importance. There is the temptation to take some shortcuts. This paper offers some suggestions for successful and meaningful sample-size determination. Also discussed is the possibility that sample size may not be the main issue, that the real goal is to design a high-quality study. Finally, criticism is made of some ill-advised shortcuts relating to power and sample size.

Key words: Power; Sample size; Observed power; Retrospective power; Study design; Cohen's effect measures; Equivalence testing;

1 Sample size and power

Statistical studies (surveys, experiments, observational studies, etc.) are always better when they are carefully planned. Good planning has many aspects. The problem should be carefully defined and operationalized. Experimental or observational units must be selected from the appropriate population. The study must be randomized correctly. The procedures must be followed carefully. Reliable instruments should be used to obtain measurements.

Finally, the study must be of adequate size, relative to the goals of the study. It must be "big enough" that an effect of such magnitude as to be of scientific significance will also be statistically significant. It is just as important, however, that the study not be "too big," where an effect of little scientific importance is nevertheless statistically detectable. Sample size is important for economic reasons: An under-sized study can be a waste of resources for not having the capability to produce useful results, while an over-sized one uses more resources than are necessary. In an experiment involving human or animal subjects, sample size is a pivotal issue for ethical reasons. An under-sized experiment exposes the subjects to potentially harmful treatments without advancing knowledge. In an over-sized experiment, an unnecessary number of subjects are exposed to a potentially harmful treatment, or are denied a potentially beneficial one.

For such an important issue, there is a surprisingly small amount of published literature. Important general references include Mace (1964), Kraemer and Thiemann (1987), Cohen (1988), Desu and Raghavarao (1990), Lipsey (1990), Shuster (1990), and Odeh and Fox (1991). There are numerous articles, especially in biostatistics journals, concerning sample-size determination for specific tests. Also of interest are studies of the extent to which sample size is adequate or inadequate in published studies; see Freiman et al. (1986) and Thornley and Adams (1998). There is a growing amount of software for sample-size determination, including *nQuery Advisor* (Elashoff, 2000), *PASS* (Hintze, 2000), *UnifyPow* (O'Brien, 1998), and *Power and Precision* (Borenstein et al., 1997). Web resources include a comprehensive list of power-analysis software (Thomas, 1998) and online calculators such as Lenth (2000). Wheeler (1974) provides some useful approximations for use in linear models; Castelloe (2000) gives an up-to-date overview of computational methods.

There are several approaches to sample size. For example, one can specify the desired width of a confidence interval and determine the sample size that achieves that goal; or a Bayesian approach can be used where we optimize some utility function—perhaps one that involves both precision of estimation and cost. One of the most popular approaches to sample-size determination involves studying the power of a test of hypothesis. It is the approach emphasized here, although much of the discussion is applicable in other contexts. The power approach involves these elements:

1. Specify a hypothesis test on a parameter θ (along with the underlying probability model for the data).

2. Specify the significance level α of the test.

3. Specify an *effect size* $\tilde{\theta}$ that reflects an alternative of scientific interest.

4. Obtain historical values or estimates of other parameters needed to compute the power function of the test.

5. Specify a target value $\tilde{\pi}$ of the power of the test when $\theta = \tilde{\theta}$.

Notationally, the power of the test is a function $\pi(\theta, n, \alpha, \ldots)$ where n is the sample size and the "$\ldots$" part refers to the additional parameters mentioned in step 4. The required sample size is the smallest integer n such that $\pi(\tilde{\theta}, n, \alpha, \ldots) \geq \tilde{\pi}$.
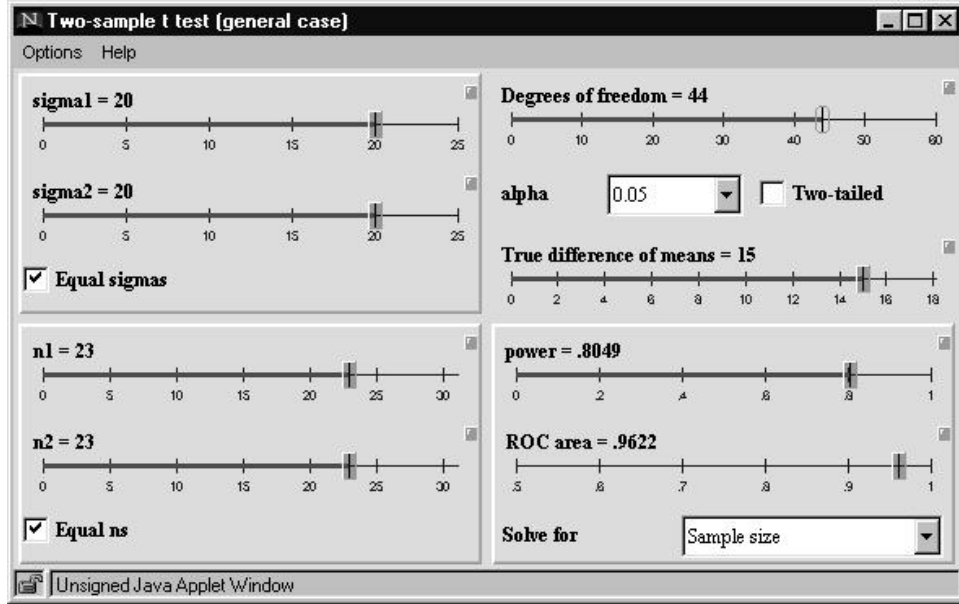
Figure 1: Software solution (Java applet in Lenth, 2000) to the sample-size problem in the blood-pressure example.

Example   To illustrate, suppose that we plan to conduct a simple two-sample experiment comparing a treatment with a control. The response variable is systolic blood pressure (SBP), measured using a standard sphygmomanometer. The treatment is supposed to reduce blood pressure; so we set up a one-sided test of $H_0 : \mu_T = \mu_C$ versus $H_1 : \mu_T < \mu_C$, where μ_T is the mean SBP for the treatment group and μ_C is the mean SBP for the control group. Here, the parameter $\theta = \mu_T - \mu_C$ is the effect being tested; thus, in the above framework we would write $H_0 : \theta = 0$ and $H_1 : \theta < 0$.

The goals of the experiment specify that we want to be able to detect a situation where the treatment mean is 15 mm Hg lower than the control group; i.e., the required effect size is $\tilde{\theta} = -15$. We specify that such an effect be detected with 80% power ($\tilde{\pi} = .80$) when the significance level is $\alpha = .05$. Past experience with similar experiments—with similar sphygmomanometers and similar subjects—suggests that the data will be approximately normally distributed with a standard deviation of $\sigma = 20$ mm Hg. We plan to use a two-sample pooled t test with equal numbers n of subjects in each group.

Now we have all of the specifications needed for determining sample size using the power approach, and their values may be entered in suitable formulas, charts, or power-analysis software. Using the computer dialog shown in Figure 1, we find that a sample size of $n = 23$ per group is needed to achieve the stated goals. The actual power is .8049.

The example shows how the pieces fit together, and that with the help of appropriate software, sample-size determination is not technically difficult. Defining the formal hypotheses and significance level are familiar topics taught in most introductory statistics courses. Deciding on the target power is less familiar. The idea is that we want to have a reasonable chance of detecting the stated effect size. A target value of .80 is fairly common and also somewhat minimal—some authors argue for higher powers such as .85 or .90. As power increases, however, the required sample size increases at an increasing rate. In this example, a target power of $\tilde{\pi} = .95$ necessitates a sample size of $n = 40$—almost 75% more than is needed for a power of .80.

The main focus of this article is the remaining specifications (items (3) and (4)). They can present some real difficulties in practice. Who told us that the goal was to detect a mean difference of 15 mm Hg?

3

How do we know that $\sigma = 20$, given that we are only planning the experiment and so no data have been collected yet? Such inputs to the sample-size problem are often hard-won, and the purpose of this article is to describe some of the common pitfalls. These pitfalls are fairly well known to practicing statisticians, and are discussed in several applications-oriented papers such as Muller and Benignus (1992) and Thomas (1997); but there is not much discussion of such issues in the "mainstream" statistical literature.

Obtaining an effect size of scientific importance requires obtaining meaningful input from the researcher(s) responsible for the study. Conversely, there are technical issues to be addressed that require the expertise of a statistician. Section 2 talks about each of their contributions. Sometimes, there are historical data that can be used to estimate variances and other parameters in the power function. If not, a pilot study is needed. In either case, one must be careful that the data are appropriate. These aspects are discussed in Section 3.

In many practical situations, the sample size is mostly or entirely based on non-statistical criteria. Section 4 offers some suggestions on how to examine such studies and help ensure that they are effective. Section 5 makes the point that not all sample-size problems are the same, nor are they all equally important. It also discusses the interplay between study design and sample size.

Since it can be so difficult to address issues such as desired effect size and error variances, people try to bypass them in various ways. One may try to redefine the problem, or rely on arbitrary standards; see Section 6. We also argue against various misguided uses of retrospective power in Section 7.

The subsequent exposition makes frequent use of terms such as "science" and "research." These are intended to be taken very broadly. Such terms refer to the acquisition of knowledge for serious purposes, whether they be advancement of a scholarly discipline, increasing the quality of a manufacturing process, or improving our government's social services.

2 Eliciting effect size

Recall that one step in the sample-size problem requires eliciting an effect size of scientific interest. It is not up to a statistical consultant to decide this; however, it is her responsibility to try to elicit this information from the researchers involved in planning the study.

The problem is that researchers often don't know how to answer the question, or don't know what is being asked, or don't recognize it as a question that they are responsible for answering. This is especially true if it is phrased too technically, e.g., "How big a difference would be important for you to be able to detect with 90% power using a Satterthwaite t test with $\alpha = .05$?" The response will likely be "Huh??" or "You're the statistician—what do you recommend?" or "Any difference at all would be important."

Better success is achieved by asking concrete questions and testing out concrete examples. A good opening question is: "What results do you expect (or hope to see)?" In many cases, the answer will be an *upper* bound on $\tilde{\theta}$. That is because the researcher probably would not be doing the study if she does not expect the results to be scientifically significant. In this way, we can establish a *lower* bound on the required sample size. To narrow it down further, ask questions like: "Would an effect of half that magnitude [but give the number] be of scientific interest?" Meanwhile, be aware that halving the value of $\tilde{\theta}$ will approximately quadruple the sample size. Trial calculations of n for various proposals will help to keep everything in focus. You can also try a selection of effect sizes and corresponding powers, e.g., "With 25 observations, you'll have a 50% chance of detecting a difference of 9.4 mm Hg, and a 90% chance of detecting a difference of 16.8 mm Hg." Along the same lines, you can show the client the gains and losses in power or detectable effect size due to increasing or decreasing n, e.g., "if you're willing to pay for 6 more subjects per treatment, you'll be able to detect a difference of 15 mm Hg with 90% power."

It may be beneficial to ask about relative differences instead of absolute ones; e.g., "Would a 10% decrease in SBP be of practical importance?" Also, it may be effective to reverse the context to what *cannot* be detected: "What is the range of clinical indifference?" And you can appeal to the researcher's values: "If

you were the patient, would the benefits of reducing SBP by 15 mm Hg outweigh the cost, inconvenience, and potential side effects of this treatment?" This latter approach is more than just a trick to elicit a response, because such value judgments are of great importance in justifying the research.

Boen and Zahn (1982), page 119–122, discuss some of the human dynamics involved in discussing sample size (mostly as distinct from effect size). They suggest asking directly for an upper bound on sample size, relating that most clients will respond readily to this question. Given the above method for establishing a lower bound, things might get settled pretty quickly—unless, of course, the lower bound exceeds the upper bound! (See Section 4 for suggestions if that happens.)

Industrial experiments offer an additional perspective for effect-size elicitation: the bottom line. Sample size relates to the cost of the experiment, and target effect size is often related directly to hoped-for cost savings due to process improvement. Thus, sample size may be determinable from a type of cost/benefit analysis.

Note that the discussion of tradeoffs between sample size and effect size requires both the technical skills of the statistician and the scientific knowledge of the researcher. Scientific goals and ethical concerns must both be addressed. The discussion of ethical values involves everyone, including researchers, statisticians, and lab technicians.

3   Finding the right variance

Power functions usually involve parameters unrelated to the hypotheses. Most notably, they often involve one or more variances. For instance, in the SBP example above, we need to know the residual variance of the measurements in the planned two-sample experiment.

Our options are to try to elicit a variance from the experimenter by appealing to his experience, to use historical data, or to conduct a pilot study. In the first approach, investigators often have been collecting similar data to that planned for some time, in a clinical mode if not in a research mode; so by talking to them in the right way, it may be possible to get a reasonable idea of the needed variance. One idea is to ask the researcher to construct a histogram showing how they expect the data to come out. Then you can apply simple rules (e.g., the central 95-percent range comprises about four standard deviations, if normal). You can ask for anecdotal information: "What is the usual range of SBPs? Tell me about some of the smallest and largest SBPs that you have seen." Discuss the stories behind some of the extreme measurements to find out to what extent they represent ordinary variations. (Such a discussion might provide additional input to the effect-size question as well.)

Historical data include data collected by the investigator in past experiments or work, and data obtained by browsing the literature. Historical or pilot data do *not* need to follow the same design as the planned study; but one must be careful that the right variance is being estimated. For example, the manufacturer of the sphygmomanometers to be used in the SBP experiment may have published test results that show that the standard deviation of the readings is 2.5 mm Hg. This figure is not appropriate for use in sample-size determination, because it probably reflects variations in readings made on the same subject under identical conditions. The residual variation in the SBP experiment includes variations among subjects.

In general, careful identification and consideration of sources of variation in past studies is much more important than that they be of the same design. In a blood-pressure-medication study, these sources include: patient attributes (sex, age, risk factors, demographics, etc.), instrumentation, how, when, and who administers medication and collects data, blind or non-blind studies, and other factors. In a simple one-factor study, suppose that we have past data on a two-factor experiment where male and female subjects were separately randomized to groups who received different exercise regimens; and that the response variable is SBP measured using instruments identical to those that you plan to use. This may provide useful data for planning the new study—but you have to be careful. For example, the residual variance of the old study

does not include variations due to sex. If the new study uses subjects of mixed sex, then the variation due to sex must be included in the error variance used in sample-size planning. Another issue is whether, in each study, the same person takes all measurements, or if it is done by several people—and whether their training is comparable. All of these factors affect the error variance. It can be a very difficult process to identify the key sources of variation in past studies, especially published ones. You are probably better off with complete information on all the particulars of a small number of past studies than with scant information on a large number of published studies.

After identifying all of the important sources of variation, it may be possible to piece together a suitable estimate of error variance using variance-component estimates. Skill in thinking carefully about sources of variation, and in estimating them, is an important reason why a statistician should be involved in sample-size planning.

There may be substantial uncertainty in variance estimates obtained from historical or pilot data (but in many cases, the fact that sample-size planning is considered *at all* is a big step forward). There is some literature on dealing with variation in pilot data; a good starting point is Taylor and Muller (1995). Also, Muller and Benignus (1992) and Thomas (1997) discuss various simpler ways of dealing with these issues, such as sensitivity analyses.

Finally, once the data are collected, it is useful to compare the variances actually observed with those that were used in the sample-size calculations. This will not help in the design of the present study, but is helpful as part of a learning process leading to better success in designing future studies. Big discrepancies should be studied to try to identify what was overlooked; small discrepancies help build a track record of success. On a related matter, careful documentation of a study and its analysis is important not only for proper reporting of the present study, but for possible use as historical data in future sample-size determinations.

4 What to do if you have no choice about sample size

Often, a study has a limited budget, and that in turn determines the sample size. Another common situation is that a researcher or senior colleague (or indeed a whole research area) may have established some convention regarding how much data is "enough." Some amusing anecdotes of the latter type are related in Boen and Zahn (1982), pages 120–121.

It is hard to argue with budgets, journal editors, and superiors. But this does not mean that there is no sample-size problem. As we discuss in more detail in Section 5, sample size is but one of several quality characteristics of a statistical study; so if n is held fixed, we simply need to focus on other aspects of study quality. For instance, given the budgeted (or imposed) sample size, we can find the effect size $\ddot{\theta}$ such that $\pi(\ddot{\theta}, n, \alpha, \ldots) = \tilde{\pi}$. Then the value of $\ddot{\theta}$ can be discussed and evaluated relative to scientific goals. If it is too large, then the study is under-powered, and then the recommendation depends on the situation. Perhaps this finding may be used to argue for a bigger budget. Perhaps a better instrument can be found that will bring the study up to a reasonable standard. Last (but definitely not least), re-consider possible improvements to the study design that will reduce the variance of the estimator of θ, e.g., using judicious stratification or blocking..

Saying that the study should not be done at all is probably an unwelcome (if not totally inappropriate) message. The best practical alternatives are to recommend that the scope of the study be narrowed (e.g., more factors are held fixed), or that it be proposed as part of a sequence of studies. The point is that just because the sample size is fixed does not mean that there are not some other things that can be changed in the design of the study.

It is even possible that $\ddot{\theta}$ (as defined above) is *smaller* than necessary—so that the planned study is over-powered. Then the size of study could be reduced, perhaps making the resources available for some other study that is less adequate. (As Boen and Zahn (1982) points out, even this may not be welcome news, due

to prejudices about what sample size is "right.") An alternative might be to keep the sample size fixed, but to broaden the scope of the study (broader demographics of subjects, additional suppliers of raw material, etc.); that will make the results more widely applicable, thus obtaining more "bang for the buck." When animal or human subjects are involved, an over-powered study raises a serious ethical dilemma. Fortunately, institutional review boards are becoming more sophisticated on power and sample-size issues, so there is hope that there will be fewer unnecessarily large studies in the future.

5   Not all sample-size problems are the same

Not all sample-size problems are the same, nor is sample size equally important in all studies. For example, the ethical issues in an opinion poll are very different from those in a medical experiment, and the consequences of an over- or under-sized study also differ.

In an industrial experiment, it may take only minutes to perform an experimental run, in which case there are few consequences if the experiment is too small. A clinical study may be relatively short-term and involve some potential risk to patients. In such situations, it may be desirable to proceed in a sequence of small experiments, with interim analyses in between.

Sample-size issues are usually more important when it takes a lot of time to collect the data. An agricultural experiment may require a whole growing season, or even a decade, to complete. If its sample size is not adequate, the consequences are severe. It thus becomes much more important to plan carefully, and to place greater emphasis on hedging for the possibility of under-estimating the error variance, since that would cause us to under-estimate the sample size.

There is a continuum of situations in between. Part of the conversation in sample-size planning should center on the consequences of getting it wrong: What if we find ourselves wanting a follow-up study? How much will that set us back? Can we budget for it? What are the ethical issues in a study that is too large or too small? Answers to such questions will help to decide how liberal or conservative we need to be in sample-size calculations.

Sample-size problems also vary widely in their complexity. If normally distributed data can be expected, and we have, say, a randomized complete-block design, then available tables, charts, or software can be used. If the analysis will be a multi-factor mixed-effects analysis of variance for balanced data, there are a number of tests to consider and a number of variance components to keep track of; but a good mathematical statistician might still be able to find or improvise a reasonable answer (unfortunately, most textbooks, if they mention sample size at all, don't go beyond the very simplest scenarios). If there will be substantial non-response, censoring, or correlated multivariate responses, the only recourse may be simulation of several plausible scenarios to get an idea of how good the proposed study is. Additional complications can hold for attribute data, due to failures of asymptotic tests, inability to achieve a stated size due to discreteness, or unusual situations such as inferences about rare attributes (Wright, 1997). Simulation methods again are helpful in addressing these problems.

Finally, there is really no such thing as just a sample-size problem. Sample size is but one aspect of study design. When you are asked to help determine sample size, a lot of questions must be asked and answered before you even get to that one: Exactly what are the goals? Are you really asking about sample size? Is it even a statistical study? What is the response variable, how do you plan to take measurements, and are there alternative instruments? What can go wrong? What is your estimate of the non-response rate? What are the important sources of variation? How can we design the study to estimate θ efficiently? What is the time frame? What are the other practical constraints? You may often end up *never* discussing sample size because these other matters override it in importance.

6   Avoid "canned" effect sizes

Most of the rest of this article discusses some practices to avoid. First and foremost of these is the all-too-common misuse of the effect-size measures described in Cohen (1988). For a pooled t test, Cohen defines an effect-size d to be the target difference of means divided by the error standard deviation (i.e., $d = \tilde{\theta}/\sigma$). I call d a *standardized* effect size because it is unit-free, compared with an absolute effect size like $\tilde{\theta}$ that has units attached (such as mm Hg). Cohen suggests guidelines for d: it is "small," "medium," or "large" if d is .20, .50, or .80 respectively. These assessments are based on an extensive survey of statistics reported in the literature in the social sciences. Accordingly, many researchers have been misguided into using these as targets; e.g., find the sample size needed to detect a "medium" effect at 80% power.

As is discussed above, eliciting meaningful effect sizes and estimating error variances constitute two potentially difficult obstacles in addressing sample-size problems. Using Cohen's effect sizes as targets, we just appeal to conventions and avoid having to talk about either $\tilde{\theta}$ or σ—sounds like a good deal, right? Wrong! Consider, for example, an industrial experiment where measurements could be made using a coordinate-measuring machine (accurate to a few microns), a vernier caliper (accurate to a few thousandths of an inch), or a school ruler (accurate to a sixteenth of an inch). No matter which you use, you get the same sample size for a "medium" effect at 80% power. Obviously, your choice of instrumentation has a huge effect on the results, and so it should affect your sample-size calculations. There is no honest way to avoid talking about $\tilde{\theta}$ and σ separately.

The combination of α, $\tilde{\pi}$, and a standardized effect size completely determines the sample size for any study of a specified design. Thus, asking for a small, medium, or large standardized effect size is just a fancy way of asking for a large, medium, or small sample size, respectively. If only a standardized effect is sought without regard for how this relates to an absolute effect, the sample-size calculation is just a pretense.

Standardized-effect-size goals are misused in many other situations. For example, in simple linear regression of a variable y on another variable x, the correlation (or squared correlation) between x and y can serve as a standardized effect-size measure. This measure encapsulates three quantities: the slope of the line, the error variance, and the variance of the x values. These are, respectively, absolute effect size, variance, and experimental design—the three aspects of study design emphasized most strongly in the preceding sections. It is mandatory that these three quantities be considered separately, rather than being lumped together into a single R^2 measure.

7   Avoid retrospective planning

Another way to dance around having to elicit effect sizes and variances is to table those issues until after the study is completed. At that point, we will have estimates of all the effect sizes and variances we need, and so are able to a kind of retrospective sample-size or power analysis.

Let us look first at the case where the statistical test is "significant"; then there is a clear result that can be acted upon (provided that it is also of sufficient magnitude to be of scientific importance). The present study will not continue, or at least the focus will shift to something else. Of course, there is the possibility that the study included far more data than were really needed, and this can be grounds for criticism on ethical or economic grounds.

On the other hand, if the test turns out to be "non-significant," the researchers may want to design a follow-up study with sufficient data so that an effect of the same size as that observed in the study would be detected. In other words, one specifies $\tilde{\theta} = \hat{\theta}$ in designing the follow-up study. This is a form of *post hoc* or retrospective effect-size elicitation. This is quite different from effect-size elicitation based on scientific goals. The goal now is really to collect enough additional data to obtain statistical significance, while ignoring scientific meaning. It is asterisk-hunting. While acknowledging that many journals seem to use

statistical significance as a yardstick to measure publishability of research results, this tendency can hardly be applauded.

There is another popular strategy, equally bad, for dealing with a nonsignificant finding: it is to attempt to make an inference based on the power at the observed effect size:

$$\pi_{obs} = \pi(\hat{\theta}, n, \alpha, \dots)$$

where $\hat{\theta}$ is the observed estimate of θ. We will refer to this quantity as the "observed power." Despite its popularity, observed power only confuses the issue; it does not provide any additional insight beyond the results of the statistical test. The automotive analogy is that if your car made it to the top of the hill, it was powerful enough; if it didn't, it wasn't powerful enough.

Hoenig and Heisey (2001) explain the pitfalls of observed power in detail. The main technical point is that it can be shown that π_{obs} is a decreasing function of the P value of the test; we already know how to interpret P values, so we don't need observed power. One of Hoenig and Heisey's most important points concerns a common claim made by proponents of observed power: that if the test is nonsignificant but the observed power is high, then there is strong statistical evidence supporting the belief that H_0 is true. However, since observed power increases as the P value decreases, high observed power constitutes evidence *against* the null hypothesis—the opposite of the proponents' claims.

I have also seen observed power used in a way that exaggerates or distorts the statistical results: "Not only is it significant, but the test is really powerful!"—or: "The results are not significant, but that's because the test is not very powerful." The relation between π_{obs} and P values shows that if the test is significant, the power is bound to be high, and when it's non-significant, the power is bound to be low. (In the case of a t test, or other statistic that has a fairly symmetric distribution, the borderline case where $P = \alpha$ corresponds to $\pi_{obs} \approx 50\%$.)

There is yet another type of retrospective power worth mentioning (and discounting). Suppose that we examine the value of $\pi(\tilde{\theta}, n, \alpha, \dots)$ after collecting the data (using the data to estimate auxiliary parameters such as the error SD). This differs from observed power in that it uses a specified effect size $\tilde{\theta}$ of scientific meaning, rather than the observed effect $\hat{\theta}$. If this retrospective power is high in a case where the null hypothesis is not rejected, it is claimed that one can establish a reasonable certainty that the effect size is no more than $\tilde{\theta}$. (It is also possible to construct confidence bounds on this power). Again, this is a faulty way to do inference; Hoenig and Heisey (2001) point out out that it is in conflict with an inference based on a confidence interval. For example, in a t-test situation, a 95% confidence interval for θ will contain $\tilde{\theta}$ values that can be refuted with nearly 97.5% power; so there are values of $\tilde{\theta}$ that the confidence procedure considers plausible that are implausible based on the power calculation. A $\tilde{\theta}$ outside the confidence interval is already refuted by a statistical test, and hence a power calculation is superfluous.

Obviously, 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; one then begins contemplating the possibility that $|\theta|$ really is small, and wants to prove it. That implies a different statistical test! The correct way to proceed is not to look at the power of the original test—for which the hypotheses are formulated inappropriately—but to do a formal test of equivalence. A test of equivalence has hypotheses $H_0 : |\theta| \geq \tilde{\theta}$ versus $H_1 : |\theta| < \tilde{\theta}$ where, as before, $\tilde{\theta}$ is an effect size deemed to be of scientific importance. A good approximate test (see Schuirmann, 1987) rejects H_0 at significance level α if the $100(1 - 2\alpha)\%$ confidence interval for θ lies entirely within the interval $(-\tilde{\theta}, +\tilde{\theta})$.

8   Conclusions

Sample-size planning is often important, and almost always difficult. It requires care in eliciting scientific objectives and in obtaining suitable quantitative information prior to the study. Successful resolution of the

sample-size problem requires the close and honest collaboration of statisticians and subject-matter experts.

One cannot avoid addressing the issues of effect-size elicitation (in absolute terms) and estimating the error variance, as difficult as these may be. Standardized effects do not translate into honest statements about study goals. Observed power adds no information to the analysis, and retrospective effect-size determination shifts attention toward obtaining asterisk-studded results independent of scientific meaning. Note that both of these retrospective methods use an estimated effect size in place of one that is determined by scientific concerns. The error in confusing these is exactly the error made when statistical significance is confused with scientific significance.

It is a practical reality that sample size is not always determined based on noble scientific goals. Then it is important to evaluate the proposed study to see if it will meet scientific standards. Various types of changes to the study can be recommended if it turns out to be over- or under-powered.

Sample-size problems are context-dependent. For example, how important it is to increase the sample size to account for such uncertainty depends on practical and ethical criteria. Moreover, sample size is not always the main issue; it is only one aspect of the the quality of a study design.

Besides the power approach discussed here, there are other respectable approaches to sample-size planning, including Bayesian ones and frequentist methods that focus on estimation rather than testing. While technically different, those approaches also require care in considering scientific goals, incorporating pilot data, ethics, and study design. Good consulting techniques have broad applicability anyway; for example, many of the ideas suggested for eliciting effect size can be easily adapted to eliciting a useful prior distribution in a Bayesian context; and conversely, good techniques for eliciting a prior might be useful in setting an effect size.

References

Boen, J. R. and Zahn, D. A. (1982), *The Human Side of Statistical Consulting*, Lifetime Learning Publications, Belmont, CA.

Borenstein, M., Rothstein, H., and Cohen, J. (1997), *Power and Precision*, Biostat, Teaneck, NJ, Software for MS-DOS systems.

Castelloe, J. (2000), "Sample Size Computations and Power Analysis with the SAS System," in *Proceedings of the Twenty-Fifth Annual SAS User's Group International Conference*, Cary, NC, SAS Institute, Inc., Paper 265-25.

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

Desu, M. M. and Raghavarao, D. (1990), *Sample Size Methodology*, Academic Press, Boston.

Elashoff, J. (2000), *nQuery Advisor Release 4.0*, Statistical Solutions, Cork, Ireland, Software for MS-DOS systems.

Freiman, J. A., Chalmers, T. C., Smith, Jr., H., and Kuebler, R. R. (1986), "The Importance of Beta, the Type II Error, and Sample Size in the Design and Interpretation of the Randomized Controlled Trial: Survey of 71 "Negative" Trials," in *Medical Uses of Statistics*, eds. J. C. Bailar III and F. Mosteller, chap. 14, pp. 289–304, NEJM Books, Waltham, Mass.

Hintze, J. (2000), *PASS 2000*, Number Cruncher Statistical Systems, Kaysville, UT, Software for MS-DOS systems.

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.

Kraemer, H. C. and Thiemann, S. (1987), *How Many Subjects? Statistical Power Analysis in Research*, Sage Publications, Newbury Park, CA.

Lenth, R. V. (2000), "Java applets for power and sample size," `http://www.stat.uiowa.edu/~rlenth/Power/`.

Lipsey, M. W. (1990), *Design Sensitivity: Statistical Power for Experimental Research*, Sage Publications, Newbury Park, CA.

Mace, A. E. (1964), *Sample-size determination*, Reinhold, New York.

Muller, K. E. and Benignus, V. A. (1992), "Increasing scientific power with statistical power," *Neurotoxicology and Teratology*, 14, 211–219.

O'Brien, R. G. (1998), `UnifyPow.sas` *Version 98.08.25*, Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation, Cleveland, OH, Available for download from `http://www.bio.ri.ccf.org/power.html`.

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

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

Shuster, J. J. (1990), *CRC Handbook of Sample Size Guidelines for Clinical Trials*, CRC Press, Boca Raton.

Taylor, D. J. and Muller, K. E. (1995), "Computing Confidence Bounds for Power and Sample Size of the General Linear Univariate Model," *The American Statistician*, 49, 43–47.

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

Thomas, L. (1998), "Statistical power analysis software," `http://www.forestry.ubc.ca/conservation/power/`.

Thornley, B. and Adams, C. (1998), "Content and quality of 2000 controlled trials in schizophrenia over 50 years," *British Medical Journal*, 317, 1181–1184.

Wheeler, R. E. (1974), "Portable Power," *Technometrics*, 16, 193–201.

Wright, T. (1997), "A simple algorithm for tighter exact upper confidence bounds with rare attributes in finite universes," *Statistics and Probability Letters*, 36, 59–67.