TWO SAMPLE-SIZE PRACTICES THAT I DON'T RECOMMEND

Russell V. Lenth, University of Iowa
Department of Statistics, Iowa City, IA 52242

Key Words: Sample size, Power, Effect size, Retrospective power

1 Introduction

Participating in this panel discussion on sample-size issues was a real pleasure, and I wish to thank everyone involved.

This article summarizes the points I emphasized in that discussion. The comments are motivated in great part by the types of questions I occasionally receive by e-mail asking for help in using some software that I have developed (Lenth, 2000a). Such interactions can be quite interesting, and many inquiries have motivated me to extend and/or modify the software. However, certain questions I receive are all too common; they are ones that relate to

1. Sample-size goals based on a standardized effect size
2. Retrospective power analysis

Questions like these are easy to ask, and easy to answer—all too easy. They reflect practices that seem to be fairly firmly established; but in my opinion, they are not consistent with good science. It is my purpose here to try to explain why.

2 Standardized effect-size goals

One common question goes like this:

> What sample size is required to detect a "medium" difference with 90% power and $\alpha = .05$?

This question refers to a convention established by Jacob Cohen (Cohen, 1988) that sets norms for "small," "medium," or "large" effects. In the context of a two-sample t test, these norms correspond to $d = .20$, $d = .50$, and $d = .80$ respectively, where

$$d = (\mu_1 - \mu_2) \,/\, \sigma$$

is the standardized difference between the two means being compared. Here, $\mu_1 - \mu_2$ is the actual difference between the means for a particular alternative hypothesis under consideration, and σ is the standard deviation of the experimental error, assumed common to the two treatments. Note that a "medium" difference is half a standard deviation.

The question above, then, sets a value of d as the criterion for the sample-size problem. Proponents of this approach claim two advantages:

1. You don't need to collect pilot data or historical data to estimate σ.
2. The standards for "large," medium," and "small" are based on an extensive survey of the published literature in the social sciences, and hence reflect realistic conventions (at least in the social sciences).

A standardized effect size such as d is very useful in that it is directly relevant to creating reasonably compact tables for determining sample size. But a d value itself has no relevance as a criterion for determining sample size. As fellow panelist Janet Elashoff puts it, you need to look at the numerator and denominator of d separately.

I offer the following example to help solidify Janet's point. Suppose that a manufacturer wants to compare the mean shrinkages of injection-molded parts made with raw materials from two suppliers. Following Cohen's convention, their goal is to be able to detect a "medium"-sized difference with a power of .9, using a two-sided test at a significance level of $\alpha = .05$. Four proposals, with estimated costs and total sample sizes N are summarized below:

Proposal	Cost	N
1	\$3,500	170
2	\$5,250	170
3	\$2,800	170
4	\$3,750	172

It appears that Proposal 3 will be selected.

Now let's take a closer look at each proposal. In the table below, we show what type of design is proposed, the instrumentation to be used, the value of σ for that instrumentation, and the detectable difference of means

(half of σ for a "medium" effect) at a power of .90.

Proposal	Design Instrument Error SD	Detectable $\lvert\mu_1 - \mu_2\rvert$
1	Indep. samples (CRD)	
\$3,500	Vernier caliper	0.35 mm
$N = 170$	$\sigma = 0.70$ mm	
2	Indep. samples (CRD)	
\$5,250	Coordinate meas. mach.	0.33 mm
$N = 170$	$\sigma = 0.66$ mm	
3	Indep. samples (CRD)	
\$2,800	6-inch school ruler	0.85 mm
$N = 170$	$\sigma = 1.9$ mm	
4	Paired (block) design	
\$3,750	Vernier caliper	0.14 mm
$N = 172$	$\sigma = 0.7 = \sqrt{.64^2 + .28^2}$	

The first three proposals all use a simple completely-randomized design. The second one has the lowest σ of the three, due to the use of high-technology instrumentation—a coordinate-measuring machine—which greatly increases the cost, but for only a modest advantage (much of the variation comes from other sources). The lowest-cost proposal has by far the worst detectable difference, due to its extremely low technology. Proposal 4 uses the same (sensible) technology as Proposal 1, but gains by far the best detectable difference using a tactic that we statisticians should always keep in the forefront of our thinking: a good experimental design. By blocking on subjects, an important source of variation is eliminated from the treatment comparison, effectively reducing the error standard deviation from .70 to .28. Proposal 4 costs just a bit more than Proposal 1 due to the slightly increased sample size associated with having fewer degrees of freedom for error. But if the engineers' goal is to be able to detect a difference of, say, 0.25 mm, we can get by with a lot less data (and lower cost).

The lesson here is that when we focus on a standardized effect size, we are ignoring many of the important issues that deserve careful consideration in designing any statistical study. We get no credits (or demerits) for using good (or bad) measurement procedures, and we actually get slight negative credit for using a good design. (A note here: A colleague points out that a more careful researcher would not be misled by the design issue. For example, a program such as *nQuery Advisor* (Elashoff, 2000) would make an adjustment to the error variance based on the anticipated intraclass correlation. However, that still does not cure the faults associated with standardized effect specifications.) It is always important to think in terms of actual, absolute effect sizes, in the same units of measurement as where the inference is to be made. There is really no honest way around addressing both the numerator and denominator of d separately.

Standardized effect sizes have been defined for a number of contexts—analysis of variance, paired t tests, etc. In a regression context, it is popular to use the correlation ρ or the coefficient of multiple determination ρ^2 as a target. Doing so is even a worse mistake than in the example above, because ρ and ρ^2 involve not only one or more absolute effect sizes (coefficients β_j) and the error variance σ^2, but also the variance(s) of the predictor(s). All three of these elements should be considered separately in designing a study and determining sample size.

3 Retrospective power

Retrospective power analysis comprises a number of different practices that involve computing the power of a test based on observed data. Personally, I think I can do without all such practices; but some of them are more understandable than others. The one that I really don't like is the idea of computing power using observed data, with the observed error variance and the observed effect size.

Again, I offer an example. Figure 1 shows the default output generated by the regression procedure in SAS Analyst (SAS Institute, Inc., 1999) with the "Perform power analysis" option checked in the "Tests" menu. (Several other statistical packages can produce similar results, presumably in response to customer demands.) The data are measurements of 202 Australian athletes (Cook and Weisberg, 1999, p. 438). The top part of the output is conventional statistical output for a regression procedure; the bottom shows retrospective power for the observed effect sizes at significance level $\alpha = .05$. Also shown are "least significant numbers" (LSNs), which reflect the sample size required to achieve significance at the observed effect size; they are truncated at 1002.

To show the folly of these retrospective power calculations, selected portions of Figure 1 are re-organized in the following table, and sorted by P value.

Source	t ratio	P value	Power	LSN
LBM	−18.02	$< .0001$	0.999	14
Wt	10.48	$< .0001$	0.999	16
SqrtSSF	9.25	$< .0001$	0.999	17
Sex	4.78	$< .0001$	0.997	38
Hg	2.25	0.0258	0.609	157
BMI	1.61	0.1098	0.359	304
Hc	0.71	0.4816	0.108	1002
Height	−0.74	0.4627	0.113	1002
WCC	−0.48	0.6306	0.077	1002
Ferr	0.44	0.6574	0.073	1002
RCC	−0.36	0.7172	0.065	1002

It is immediately obvious that as the P value increases, retrospective power decreases, and least significant number increases. In fact, both are simply transformations of the P values. It can further be shown that when the P value is equal to α, the retrospective power is approximately 0.5. That is true because the empirical effect size is right at the boundary of the critical region, so that about half of the probability falls in the critical region.

There is simple intuition behind results like these: If my car made it to the top of the hill, then it is powerful enough to climb that hill; if it didn't, then it obviously isn't powerful enough. Retrospective power is an obvious answer to a rather uninteresting question. A more meaningful question is to ask whether the car is powerful enough to climb a particular hill never climbed before; or whether a different car can climb that new hill. Such questions are prospective, not retrospective.

The fact that retrospective power adds no new information is harmless in its own right. However, in typical practice, it is used to exaggerate the validity of a significant result ("not only is it significant, but the test is really powerful!"), or to make excuses for a nonsignificant one ("well, P is .38, but that's only because the test isn't very powerful"). The latter case is like blaming the messenger.

Similarly, LSNs don't add new information. True, if we collect more data to bring N up to the LSN, and the effect size stays the same, then we'll obtain statistical significance. Such a strategy is strictly asterisk-hunting: let's do whatever it takes to make $P < .05$. Instead, as in the preceding section, I recommend consulting with subject-matter experts before the data are collected to determine absolute effect-size goals that are consistent with the scientific goals of the study.

For further discussion of retrospective power from a scientist's perspective, I recommend Thomas (1997).

4 Discussion

Sample-size determination is serious and important business. It is the one place in the process of collecting and analyzing data where scientific goals can be addressed. That takes hard work and careful thinking. The practices I criticize in this article are popular primarily because they are easy, and they are easy because they bypass the detailed study that is really necessary to do it right.

This article contains selected topics (plus a few embellishments) from a longer report on sample-size practices (Lenth, 2000b). That report also discusses some of the consulting aspects of sample-size determination, what to do when the sample size is fixed, how to estimate σ, and the fact that not all sample-size problems are the same.

References

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

Cook, R. D. and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics*, Wiley, New York.

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

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

Lenth, R. V. (2000b), "Seven habits of highly effective sample-size determination," submitted.

SAS Institute, Inc. (1999), "SAS Proprietary Software," Version 8.

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

Figure 1: Retrospective power-analysis example.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-8.62641	6.15867	-1.40	0.1629
BMI	1	0.22249	0.13850	1.61	0.1098
Ferr	1	0.00052053	0.00117	0.44	0.6574
Hc	1	0.03713	0.05266	0.71	0.4816
Hg	1	-0.09203	0.12505	-0.74	0.4627
Height	1	0.07641	0.03400	2.25	0.0258
LBM	1	-0.82153	0.04559	-18.02	<.0001
RCC	1	-0.10165	0.28024	-0.36	0.7172
Sex	1	1.08958	0.22775	4.78	<.0001
SqrtSSF	1	1.04342	0.11279	9.25	<.0001
WCC	1	-0.01348	0.02798	-0.48	0.6306
Wt	1	0.63786	0.06084	10.48	<.0001

Power Analysis

Dependent Variable	Source	Sum of Squares Type	Alpha	Power	Least Significant Number
PctBodyFat	BMI	Type II	0.05	0.359	304
PctBodyFat	Ferr	Type II	0.05	0.073	1002
PctBodyFat	Hc	Type II	0.05	0.108	1002
PctBodyFat	Height	Type II	0.05	0.609	157
PctBodyFat	Hg	Type II	0.05	0.113	1002
PctBodyFat	LBM	Type II	0.05	0.999	14
PctBodyFat	RCC	Type II	0.05	0.065	1002
PctBodyFat	Sex	Type II	0.05	0.997	38
PctBodyFat	SqrtSSF	Type II	0.05	0.999	17
PctBodyFat	WCC	Type II	0.05	0.077	1002
PctBodyFat	Wt	Type II	0.05	0.999	16