Neurotoxicology and Teratology, Vol. 14, pp. 211-219, 1992
Printed in the U.S.A. All rights reserved.

0892-0362/92 $5.00 + .00
Copyright © 1992 Pergamon Press Ltd.

Increasing Scientific Power With Statistical Power

KEITH E. MULLER[1]* AND VERNON A. BENIGNUS†

**Department of Biostatistics, CB#7400, †Human Studies Division, HERL, US EPA, and Department of Psychology, University of North Carolina, Chapel Hill, NC 27599*

Received May 1991; Accepted September 1991

MULLER, K. E. AND V. A. BENIGNUS. *Increasing scientific power with statistical power.* NEUROTOXICOL. TERATOL. 14(3), 211-219, 1992.—A survey of basic ideas in statistical power analysis demonstrates the advantages and ease of using power analysis throughout the design, analysis, and interpretation of research. The power of a statistical test is the probability of rejecting the null hypothesis of the test. The traditional approach to power involves computation of only a single power value. The more general power curve allows examining the range of power determinants, which are sample size, population difference, and error variance, in traditional ANOVA. Power analysis can be useful not only in study planning, but also in the evaluation of existing research. An important application is in concluding that no scientifically important treatment difference exists. Choosing an appropriate power depends on: a) opportunity costs, b) ethical trade-offs, c) the size of effect considered important, d) the uncertainty of parameter estimates, and e) the analyst's preferences. Although precise rules seem inappropriate, several guidelines are defensible. First, the sensitivity of the power curve to particular characteristics of the study, such as the error variance, should be examined in any power analysis. Second, just as a small type I error rate should be demonstrated in order to declare a difference nonzero, a small type II error should be demonstrated in order to declare a difference zero. Third, when ethical and opportunity costs do not preclude it, power should be at least .84, and preferably greater than .90.

Choosing sample size   Research design   Finding no effect   Meta-analysis   Sensitivity analysis

INTRODUCTION

Motivation

TOXICOLOGISTS and teratologists spend much of their time and other resources on the design, conduct, analysis, and reporting of research. Statistical methods for testing hypotheses dominate current practice. Any increase in the efficiency of a design, analysis, or the planning process would benefit scientists. Statistical power analysis can enable a scientist to achieve such efficiencies.

The intent here is to encourage the use of power analysis by explaining the advantages of doing so. Studies with either inadequate or excessive sensitivity waste scientists' time and resources. Considering risk of possible harm to subjects, ethical concerns strengthen the desire for optimal sample size. Statistical power analysis substantially enhances a scientist's ability to plan and evaluate research. Readily available and inexpensive computer programs have made using power analysis convenient.

The emphases in this article are on a) a general conceptual approach, b) the reasons for conducting a power analysis, and c) the range of applications. The mechanics of how to conduct power analysis for particular applications are not covered. It is hoped that the references will allow the readers to acquire the understanding of the techniques for their particular applications. The actual conduct of such calculations is usually best left to a computer program.

Analysis of variance (ANOVA), especially involving repeated measures, is probably the most widely used data analysis method in toxicology and teratology. The widespread use of such methods makes them a good prototype for the following discussion. The traditional approach depends on assuming errors that are homogeneous, additive, and Gaussian. See Kirk (18) or Maxwell and Delaney (24) for comprehensive treatments. ANOVA models are special cases of a more general model often referred to as the *general linear multivariate model* (25,34). A compact and excellent introduction to the important special case of repeated measures ANOVA was provided by O'Brien and Kaiser (30). Also note that multiple regression is a special case of the general model (19). Although

[1] Requests for reprints should be addressed to Keith E. Muller, Department of Biostatistics, CB#7400, University of North Carolina, Chapel Hill, NC 27599.

Although the research described in this article has been supported by the United States Environmental Protection Agency, it has not been subjected to Agency review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

treatment levels (effects) in the model are assumed to be fixed, a useful range of random effect models can be treated, especially those resulting from repeated measures designs. Many different classes of random effect models exist which require separate and special treatment for both hypothesis testing and power analysis.

For purposes of illustrating power properties, it is sufficient to consider the simplest case of the one-way ANOVA, the two sample t test. Its properties generalize quite easily to all forms of ANOVA, as well as all other forms of the general linear multivariate model. In turn, the power properties of the general multivariate model translate to most other common data analysis methods.

The Traditional Introduction to Hypothesis Testing

To appreciate the value of statistical power analysis, it is necessary to have a clear understanding of the concept and its properties. In turn, because power is a property of a statistical hypothesis test, the principles of hypothesis testing are reviewed to provide the necessary background.

The most widespread implementation of hypothesis testing may be referred to as a Neyman–Pearson approach (named after the statisticians who first championed it). The topic is often introduced in the following way (6,19). It is assumed that one of two hypotheses, the null or the alternative, must hold. In the context of toxicology the null hypothesis may be stated as:

$$H_0\text{: toxicity of compound (under conditions tested)} = 0 \quad (1)$$

and the alternative as

$$H_A\text{: toxicity of compound (under conditions tested)} \neq 0. \quad (2)$$

For convenience of exposition, the phrase "under conditions tested" will not be explicitly stated in the remaining discussion, although it should be recognized as necessary and implied. If one rejects the null hypothesis and says a compound is toxic when it is not, then one has committed a Type I error, a false positive. The probability of a Type I error is usually denoted α. Symmetrically, if one fails to reject the null hypothesis and says a compound is not toxic when it is, then one has committed a Type II error, a false negative. The probability of a Type II error is usually denoted by β.

The assumption of two possible states of nature (toxic or not toxic), coupled to the possibility of making one of two decisions (toxic or not toxic), leads to four possible decision outcomes from testing a statistical hypothesis. Table 1 illustrates the relationships among four possible outcomes and the related probabilities. For purposes of defining the characteristics of power analysis, all characteristics of the test, including sample size, population difference, and α, are taken to be known values. For a particular choice of such values a test may be evaluated in terms of the Type I and Type II error rates, α and β. Traditionally the power of a test is defined as Power = $(1 - \beta)$, the probability of correctly detecting the population difference.

TABLE 1

TRADITIONAL DESCRIPTION OF POSSIBLE DECISIONS AND RELATED PROBABILITIES FOR A STATISTICAL HYPOTHESIS TEST

	Decision	
"Truth"	H_0: Not Toxic	H_A: Toxic
H_0: Not Toxic	Pr(Correct Negative) = $(1 - \alpha)$	Pr(False Positive) = Pr(Type I error) = α
H_A: Toxic	Pr(False Negative) = Pr(Type II error) = β	Pr(Correct Positive) = $(1 - \beta)$ = Power

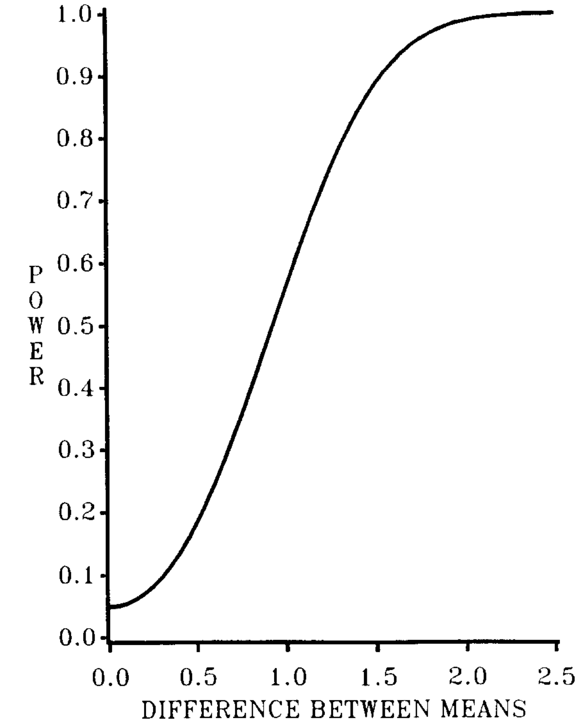

FIG. 1. Power of the test of equality of means as a function of $\delta = (\mu_1 - \mu_2)$ for two independent samples with $\alpha = .05$, $N_1 = N_2 = 10$, and $\sigma^2 = 1.0$.

The two-sample t test provides a useful example. The test is based on the assumption of independent, Gaussian errors with equal variances. To compute the probabilities in Table 1, it is necessary to specify α, group sample sizes (N_1, N_2), mean difference ($\mu_1 - \mu_2 = \delta$), and error variance (σ^2). These properties determine the value of β. If one accepts the Neyman–Pearson framework, then the quality of study design and data analysis depend on simultaneously minimizing the Type I and II error rates.

A More General Approach

The aforementioned introduction to statistical power severely limits understanding and use of the methods. A more general approach leads to defining power as the probability of saying a difference exists and considering the probability for a range of alternative hypotheses. For examples of treating power as a function, not a single value, see (7,12,13). Figure 1 is an illustration of the power (on the vertical axis) of a two-sample t test as a function of possible differences between the means, $\delta = (\mu_1 - \mu_2)$. A two-tailed test is assumed, with $\alpha = .05$, $N_1 = N_2 = 10$, and $\sigma^2 = 1.0$. The horizontal axis in Fig. 1 may be interpreted as $|\delta|$ in order to include all possible mean differences.

Figure 1, as well as all other figures and tables, will be based on this same situation. The numbers did not arise as data from a particular experiment. Choosing $\sigma^2 = 1.0$ leads to δ being in standard deviation units. Consequently the results can be applied to any particular experiment by appropriate interpretation of the units. For example, the horizontal axis of Fig. 1 may be taken to be $(\mu_1 - \mu_2)/\sigma$ or $|\mu_1 - \mu_2|/\sigma$.

The curve has several prototypic characteristics, including the fact that it increases monotonically as a function of size of mean difference and is an ogive. Furthermore, the intercept of the vertical axis is α, whereas the curve has an upper asymptote of 1.0. The lower bound corresponds to a power (a probability of saying a difference) of α if $\delta = 0$. The upper asymptote indicates that as δ becomes large the likelihood of detecting the treatment difference approaches certainty. The shape leads to the definition of three distinct regions in which power is low and changing slowly, or moderately low to moderately high and changing rapidly, or high and changing slowly.

The power curve for a one-tailed test asymptotes to zero as δ goes to $-\infty$, whereas the power curve for a two-tailed test is symmetric about $\delta = 0$. One-tailed tests do not exist for any hypothesis involving more than a single parameter, such as the overall test of equality of means in a one-way ANOVA with three or more groups. Even when applicable, many data analysts object to the logical properties of such tests. Hence one-tailed tests will not be considered further due to their limited utility.

Table 1 embodies the traditional introduction to power, whereas Fig. 1 embodies the concept arising from a complete description of power. In the traditional introduction power is a single point and power for a zero difference is not mentioned. The general approach allows a unified treatment of the many values of the power function.

DESCRIBING POWER

Variables Affecting Power

The two-sample t test provides a simple example for the discussion of the determinants of power. The test is based on the assumption of independent, Gaussian errors with equal variances. In general, the power of the t test depends only on α, sample size $N = (N_1 + N_2)$, the ratio of group sample sizes (N_1/N_2), mean difference $(\mu_1 - \mu_2 = \delta)$, and error variance (σ^2). Essentially the same description holds for the one-sample and paired-data t tests, with appropriate interpretation of the parameters. For example, for the paired-data test, σ^2 is the variance of the difference scores.

Usually unequal sample sizes are not considered in planning experimental research because for a fixed sample size of $N = (N_1 + N_2)$ the maximum power occurs with $N_1 = N_2 = N/2$. In practice some data are often lost due to equipment failure or human error. Additional data should never be eliminated in order to create equal sample sizes. Power is always maximized by retaining all observations, even though the resulting group sample sizes are unequal. For example, the power of a t test with group sizes of N_1 and $(N_1 + 1)$ is greater than the power of a t test with group sizes of N_1 and N_1.

The aforementioned description generalizes easily to encompass the general liner multivariate model, again assuming fixed effects and independent, homogeneous, Gaussian errors. The value of this observation lies in the fact that a large range of repeated measures models can be treated as special cases of the model. For many readers, repeated measures ANOVA is the most common data analysis. As with the t test, for a fixed α, power depends on only three components: sample size, pattern of mean differences being tested, and variance. The pattern of differences being tested, in turn, is determined by the choice of design (including covariate values), hypothesis tested, and population means. *Covariate* will be used to refer to any continuous predictor variable. Error variance must be generalized to describe the covariance matrix among any repeated measures (or multiple responses). As sample size increases, or differences being tested increase, or error variance decreases, power increases. Changes in patterns of correlations among repeated measures can either increase or decrease power.

The dependence of power for linear models on the design, hypothesis tested, and population means is illustrated in the following example. Consider planning an analysis of body weight for rats at 90 days of age. Assume that all animals will be dosed with either 0, 1, or 2 mg/kg of a substance once at age 30 days. Some basic candidate design and analysis combinations include: a) a t test based on doses of 0 and 2 mg/kg, with half of the animals in each group, b) a one-way ANOVA involving all three dose levels, c) using day 30 weight as a covariate assuming equal slopes for each dose group (traditional ANOVA with a covariate, ANCOVA), d) using day 30 weight as a covariate assuming unequal slopes for each dose group, and e) a dose × gender factorial ANOVA. The textbooks (18,24) are useful sources for explanations of the practical and technical distinctions among such designs. Taking repeated measures, which would introduce *Time* as a design factor, is very common in such research. Combining features of the alternate designs might also be attractive. Comparing two or more designs leads to many possible differences in power. If the dose-response function is monotone, then the t test will have more power than the three-group ANOVA overall test. The power difference in favor of the t test may be substantial. Except for extremely small sample sizes, comparing dose 0 to dose 2 in the ANOVA, conducting the test at nominal α, yields essentially the same power as for the t test. Adding covariates to the ANOVA or allowing unequal slopes in the ANCOVA will increase power if the gain due to reduction in error variance exceeds the loss due to reduced error degrees of freedom.

The scientist's desire to control Type I error has an important effect on power. For example, some of the analysis decisions just discussed can be explored in the data of interest. The disadvantage of such a strategy is the attendant increase in Type I error rate. Strategies for balancing Type I and Type II error rates were discussed in (29). The use of power analysis was recommended in study planning to focus the analysis plan on a small number of alternative models. For example, one might choose to use the data to test whether unequal slopes are required, then reduce to the simpler model if the test is not significant. A single planned test allows branching to the best analysis, while still allowing good control of Type I error. The reader is urged to consider the texts and articles referenced in this article for a more thorough evaluation of alternative analyses and strategies.

Reporting a Power Analysis

Rather than treat the complexity that results from considering all variables that affect power, it is usually convenient to consider only a few. Often, for example, only one choice of α, study design (not including N), and hypothesis tested are evaluated. Hence some combination of the size of the sample, hypothesized error variance, and hypothesized pattern of mean differences are varied. Table 2 provides a summary of such an analysis for a two-sample *t* test. The various values in the table constitute a sensitivity analysis, which is strongly recommended for any power analysis. The nonlinearity of the power function usually dictates varying the parameters, such as variance, mean difference, and sample size, on a logarithmic scale. A variety of values in a table such as Table 2 can greatly aid in planning a study. The values chosen can be based on quantiles of the sampling distribution of sample estimates of the parameters. The technique is illustrated in the last paragraph of this section. In addition to the impact of varying sample size, the rewards for reducing error variance or increasing the treatment difference can easily be grasped by tabling results for alternate values.

Table 3 provides an interesting contrast to Table 2 because only α was changed, from 0.05 to 0.01. Taken together, the two tables would allow a scientist the ability to evaluate the consequences of adding four variables and an associated Bonferroni correction to a planned analysis. The nonlinearity of the function leads to small differences with some combinations of parameters, and large differences with others. Such a comparison helps quantify the trade-offs in deciding how many measures to collect.

TABLE 2

TWO-SAMPLE *t* TEST POWER (×100) FOR α = .05 AS A FUNCTION OF ERROR VARIANCE (σ^2), MEAN DIFFERENCE ($\mu_1 - \mu_2 = \delta$), AND SAMPLE SIZE ($N_1 = N_2 = N/2$)

σ^2	δ	N	Power
0.32	0.5	10	23
0.32	0.5	20	46
0.32	0.5	40	78
0.32	1.0	10	69
0.32	1.0	20	96
0.32	1.0	40	>99
0.32	2.0	10	>99
0.32	2.0	20	>99
0.32	2.0	40	>99
1.00	0.5	10	11
1.00	0.5	20	19
1.00	0.5	40	34
1.00	1.0	10	29
1.00	1.0	20	56
1.00	1.0	40	87
1.00	2.0	10	79
1.00	2.0	20	99
1.00	2.0	40	>99
2.05	0.5	10	08
2.05	0.5	20	11
2.05	0.5	40	19
2.05	1.0	10	16
2.05	1.0	20	32
2.05	1.0	40	58
2.05	2.0	10	49
2.05	2.0	20	84
2.05	2.0	40	99

TABLE 3

TWO-SAMPLE *t* TEST POWER (×100) FOR α = .01 AS A FUNCTION OF ERROR VARIANCE (σ^2), MEAN DIFFERENCE ($\mu_1 - \mu_2 = \delta$), AND SAMPLE SIZE ($N_1 = N_2 = N/2$)

σ^2	δ	N	Power
0.32	0.5	10	07
0.32	0.5	20	22
0.32	0.5	40	54
0.32	1.0	10	37
0.32	1.0	20	84
0.32	1.0	40	>99
0.32	2.0	10	96
0.32	2.0	20	>99
0.32	2.0	40	>99
1.00	0.5	10	03
1.00	0.5	20	06
1.00	0.5	40	14
1.00	1.0	10	10
1.00	1.0	20	29
1.00	1.0	40	67
1.00	2.0	10	48
1.00	2.0	20	93
1.00	2.0	40	>99
2.05	0.5	10	02
2.05	0.5	20	03
2.05	0.5	40	06
2.05	1.0	10	05
2.05	1.0	20	12
2.05	1.0	40	32
2.05	2.0	10	21
2.05	2.0	20	60
2.05	2.0	40	95

Figure 2 comprises a presentation of all of the information in Table 2, as well as much more information. Each function depicts the dependence of power on mean difference (δ). The three curves correspond to the three particular values of error variance treated in Table 2 (0.32, 1.00, and 2.05), with more variance generating less power. In addition to providing far more power values, a graph of the power function also compels the viewer to recognize the three distinct regions of the power curve, and to recognize the impact of uncertainty about parameters.

When the error variance value is an estimate from a pilot study, one may choose to use scale factors corresponding to the endpoints of the confidence interval around the estimate. For the example, it was assumed that $\hat{\sigma}^2$ was based on 10 error degrees of freedom, a typical value for a pilot study. For observations that can be assumed to be independent and follow a common Gaussian distribution, then $\hat{\sigma}^2$ is proportional to a χ^2. The 2.5% and 97.5% critical values for a χ^2 with 10 degrees of freedom are approximately 3.2 and 20.5. In turn the corresponding 95% confidence-interval endpoints for $\hat{\sigma}^2 = 1.00$ are $(1.00) \cdot (3.2/10) = .32$ and $(1.00) \cdot (20.5/10) = 2.05$.

PROSPECTIVE POWER ANALYSIS

One of the most common questions asked of a statistician is "How many subjects do I need?" The traditional introduc-

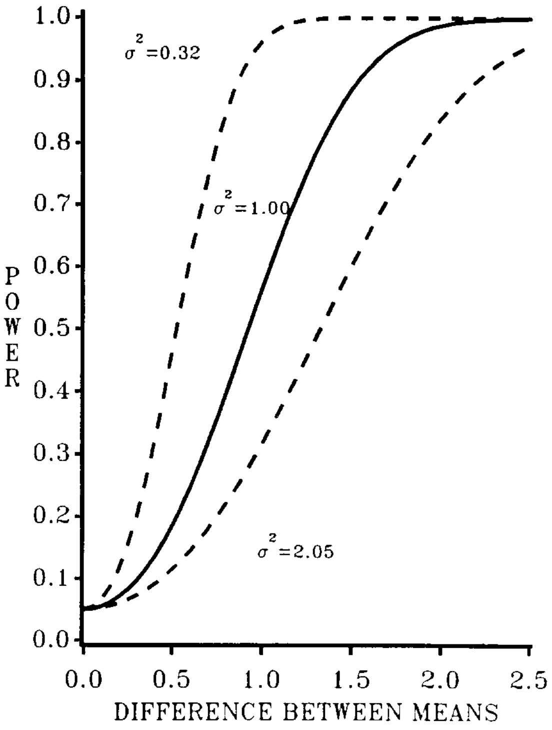

FIG. 2. Power of the test of equality of means as a function of $\delta = (\mu_1 - \mu_2)$ for two independent samples with $\alpha = .05$, $N_1 = N_2 = 10$, σ^2 either .32, 1.00, or 2.05.

tion to power analysis often leads the scientist into considering varying only sample size and ignoring the roles of mean differences and variance. More generally, power analysis allows comparing the consequences of any change in the design or analysis of a study. Such analysis is prospective in that it is conducted before the study is begun. Examples include comparing a) two-group and three-group designs, b) alternate hypothesis tests such as trend tests and pair-wise contrasts, c) significance tests evaluated with and without a Bonferroni correction, and d) even alternate test statistics such as Wilks' test and the Geisser–Greenhouse test for repeated measures.

Strategies for insuring appropriate statistical practice in toxicology were recommended by Muller et al. (29). Power analysis constitutes an integral part of their recommendations. They recommended using top-down planning to choose a well-focused, practical design. They also recommended maximizing efficiency by balancing Type I (false positive) and Type II (false negative) errors. This can be done by using appropriate power analysis and the best available statistical analysis. Their final recommendation was to implement multiple-study strategies to meet both confirmatory and exploratory goals. The results of an earlier study can be used to provide excellent estimates of the inputs to a power analysis. In this fashion, power analysis can be used in conducting and quantifying the iterative refinement inherent to the scientific process.

QUANTITATIVE METHODS FOR EVALUATING EXISTING RESEARCH

Overview

Scientists often face the need to evaluate existing research, such as in trying to synthesize a collection of research findings, or planning a new study. Furthermore, toxicologists and teratologists are increasingly asked by regulators to provide quantitative, rather than qualitative guidance. Retrospective power analysis is strongly recommended in such situations. Power analysis may be the only analysis needed, or may be used in combination with other techniques.

Meta-Analysis

Meta-analysis (9) has become popular for evaluating existing research. Several authors have provided book-length treatment of the topic (10,15,16,33). Some (5,22) considered the topic in the more general context of reviewing research. The review (32) of one book (16) contains a discussion of some technical issues of the kind that permeate the work in the area. Statisticians continue to develop the methods to formulate guidelines for their use.

Among the collection of tools available for reviewing research, meta-analysis may be likened to a chainsaw: allowing rapid but rough cuts, and dangerous to use without training. In our opinion, meta-analysis should be conducted only when re-analysis combined with power analysis is not practical.

Re-Analysis of Existing Studies

The re-analysis of combined data can be superior to meta-analysis for examining a collection of existing research. If all studies of interest use variables in the same response metric (or ones that can be mapped into a common metric), then re-analysis may be more sensitive, provide more capabilities, and be model based. For an example use of the technique see Benignus et al. (2).

Enthusiasm concerning the concept of meta-analysis should be tempered with a recognition of its limitations. Comparison to re-analysis allows highlighting the problems. As defined by the names, the inputs to meta-analysis are analyses, whereas the inputs to re-analysis are data. Meta-analysis usually operates on a single number, such as the p value or standardized mean difference (δ/σ), from each study. Hence, the meta-analyst assumes that the information bandwidth for each study is only one number wide. In contrast, the re-analyst may choose the most appropriate collection of statistics as the sufficient set.

The major barrier to re-analysis is data availability. Many scientists choose not to archive research data, at least not in any documented and machine readable format. Scientists are understandably leery about sharing data, fearing misrepresentation or being "scooped." Hence, data are rarely combined to be re-analyzed. For re-analysis to become common, scientists will need to develop and abide by ethical standards about sharing data.

Retrospective Power Analysis for Claiming "No Effect"

Power analysis is an excellent tool for the quantitative evaluation of existing research. One important application lies in interpreting a nonsignificant data analysis. Such an analysis may be characterized as retrospective power analysis because it is conducted after completion of the study and planned data analysis.

Statistical tests with nonsignificant results are particularly important in testing toxicity or teratogenicity. Defensible interpretation of such results depends on an appreciation of the distinction between statistical significance and scientific importance, as well as accurate information about the sensitivity of the study to treatment differences of scientific importance. Power analysis is a quantitative evaluation of the sensitivity of the study. Naturally the power analysis should consider treatment differences of scientifically interesting magnitude, rather than trivially small ones or absurdly large ones. It may be argued that just as a small α (Type I error) is required to declare a difference nonzero, a small β (Type II error) should be required to declare a difference zero.

We strongly endorse the position just stated. The recommendation was followed in three earlier articles. Consider (14), in which no significant differences were found in evoked potentials or reaction times due to carbon monoxide (CO) exposure. Power curves were reported to demonstrate that substantial power had been available for scientifically important differences (see Fig. 3 and surrounding discussion in 14). Also consider (1) in which no significant difference was found in symptom reporting between subjects who were exposed to high levels of CO and those who were not. Power values were provided for detecting differences in reporting rates to bolster the claim of no substantial difference (see Table 4 and surrounding discussion in 1). Finally consider (3), in which retrospective power analysis was used to plan and interpret a replication study (see the last paragraph of the Method section as well as the Discussion in 3).

Retrospective Power Analysis for Planning a Replication

A second important application of retrospective power analysis lies in planning a replication study. Such an analysis is a prospective power analysis for the replication study. The distinctive feature is that the same results comprise a retrospective power analysis for the study to be replicated. On first consideration, using power analysis in planning a replication study would seem to be a waste of time. However, simply trying to duplicate the earlier study may lead to a study with either inadequate or excessive power. Inadequate power may arise due to failure to account for uncertainty in the parameter estimates. Assuming that the variance is no greater than the previously observed value may lead to estimating power that is too large. Such a mistake would occur in roughly half of all cases unless an appropriate sensitivity analysis is conducted. For an example of a sensitivity analysis, see the earlier discussion of Table 2. More generally, inadequate power may arise because the original study had low power (even though chance favored the scientist). Similarly, excessive power may arise because the original study had very high power.

CRITERIA FOR JUDGING POWER TO BE ADEQUATE (HOW BIG IS BIG?)

Issues

In both prospective and retrospective power analysis one must specify the level of power deemed adequate. Five issues must be addressed:

1. opportunity costs,
2. ethical trade-offs,
3. the size of effect considered important,
4. the uncertainty of parameter estimates, and
5. the analyst's preference for amount of power.

The following discussion is based on the assumption that the vast diversity of applications precludes universal rules. We think that the scientific goals for a power analysis should determine the power level required. Considering each of the five issues will aid in the determination.

Opportunity Costs

Collecting any scientific observation has some cost associated with it. The costs may only be monetary but usually include time and effort. In turn, time allocated to a particular study cannot be spent on another study. Hence power analysis can be an aid to professional success by helping avoid designs with low power or wastefully high power.

If two design alternatives have equivalent power, then one may choose the alternative with lower opportunity costs. For example, a common trade-off involves collecting fewer repeated measures on a sampling unit (such as fewer pups per litter) to observe more sampling units (more litters). The scientific goals, the nature of the experimental procedure, and the facilities available to the scientist help determine which alternative should be chosen in a particular situation.

Ethical Trade-offs

Study design and data analysis usually involve many ethical issues. Research involving humans or animals has been the focus of the most discussion in this context. Current standards require basing the decision to proceed on the ratio of benefit to risk. Institutional review board members consider the risks and benefits to subjects, the scientists, and the society in judging a particular study. Some of the conflicts that result can be quantified by evaluating the power of candidate studies.

Consider the following example. In evaluating a compound with toxicity of practical importance, inadequate power favors approval of the compound, whereas excessive power involves unnecessary risk to subjects. If the compound produces no toxicity of practical importance, then inadequate power has no immediate cost, whereas excessive power still involves unnecessary risk to subjects. A delayed cost of inadequate power can occur if the original scientist or another scientist recognizes the inadequate power and subsequently conducts another study. In that situation, subjects in the first study experienced unnecessary risks. Achieving adequate power by using a high Type I error rate introduces the costs associated with falsely indicting an innocuous compound, thereby further complicating the trade-offs.

A contrasting range of examples comes from the evaluation of the efficacy of medicinal treatments. Explicit treatment of sample size consideration as an ethical issue dominates the design of clinical trials with human subjects. In evaluating a compound with efficacy of practical importance, inadequate power favors disapproval of the compound, and hence, risk to people left untreated. Excessive power involves unnecessary delay in approval (and risk to people via delayed treatment). If the compound produces no efficacy of practical importance, then inadequate power has no immediate cost, but may lead to delayed costs from conducting a second study of adequate power. Achieving adequate power by using a high Type I error rate introduces the costs associated with falsely commending a compound of medicinal value, thereby further complicating the trade-offs.

What Is a Big Effect?

The size of the effect of interest must be defined in order to compute power. In our opinion, the scientific context should be used to define the size of a minimally important difference. Even for the same response variable and for the same target population, the size of the scientifically important difference can vary across studies. For example, a particular deleterious effect observed as a consequence of chronic exposure might be trivial, while the same effect observed as a consequence of acute exposure may be substantial.

Many approaches have been taken to define an effect of consequence. One approach involves defining the importance of an effect by referring to sources of natural variation. Consider the example of figure-eight maze activity in rats. Differences between genders, times of day, and estrous states (for females) provide natural definitions of substantial effects. From the perspective of psychometric theory, this corresponds to criterion-referenced evaluation.

In contrast, norm-referenced evaluation would involve defining a big effect in terms of the SD of the response variable. This provides a second approach to defining an effect of consequence. Cohen (4) supported this approach by discussing effects as small, medium, or large in terms of SD units. The great attraction of the method, its lack of dependence on the application, may be considered to be its greatest weakness. Because the SD varies substantially as a function of experimental conditions the same absolute effect may be judged as trivial or substantial by different scientists. The same failing can occur with meta-analysis methods which often use this standardized-difference approach. Heterogeneity of experimental methods can mask or simulate heterogeneity of effects. Hence one could draw the wrong conclusion, in either direction, depending on the nature of effects and range of experimental methods.

Uncertainty of Estimates

In the calculation of power, a data analyst often must estimate or guess values for some parameters. For the example of the *t* test discussed earlier, the error variance often falls into this category. Any uncertainty in the value of such nuisance parameters introduces uncertainty in the power calculation. The results of a sensitivity analysis provide the information needed to help protect against the uncertainty. The analyst faces another compromise in balancing the uncertainty against ethical and opportunity costs.

Two methods seem practical for including the uncertainty in the power calculations. One method involves examining the data generation process and deciding what range of parameters could plausibly exist. For example, a plausible range for body temperature provides a rather small range of plausible SDs. A second method depends on having choices of parameter values which are estimates computed from a sample of data. Traditional statistical theory can be used to erect confidence limits about the estimates. In turn, the confidence limits are used in power calculations to estimate the limits on the power. The validity of the method depends on both the test and parameters of interest. Statisticians have just begun to recognize the need and to develop methods for providing accurate confidence limits for estimated power.

The aforementioned *t* test results provide a useful example of some technical issues in finding confidence limits for estimated power. Using the χ^2 distribution to create a confidence interval around the error variance, and in turn around the power value, gives results which are analytically defensible. However, simulation results demonstrate that the confidence interval on the estimated power is somewhat too small, especially with small sample estimates of variance. Some improvement may be achieved by recognizing that an estimated noncentrality parameter follows a distribution of a random variable proportional to an *F* statistic. This still does not yield an exact result for the distribution of estimated power. Furthermore estimating various combinations of parameters may have varying effects on the confidence interval for estimated power. These issues must be resolved by future statistical research. Given the current state of knowledge, the χ^2 technique may be useful in a sensitivity analysis.

How Much Power Is Enough?

Consulting statisticians often have clients ask "How many subjects do I need?" We subscribe to the position that the scientist has the responsibility to make that decision. Power analysis results typically will be evaluated in light of a) opportunity costs, b) ethical trade-offs, c) the size of effect considered important, d) the uncertainty of parameter estimates, and e) the analyst's preference for amount of power. This is an example of a general philosophy which may be described as *situational design*.

Three alternate positions will be examined concerning an appropriate level of power. Each position corresponds to a particular region or subregion of the power curve. As discussed earlier and shown in Figs. 1 and 2, the power curve for standard statistical tests includes three regions: the left shelf, the slope, and the right shelf. The left shelf region always corresponds to very low power, the slope region includes both low and high power values, and the right shelf includes only high power values. It seems highly unlikely that any scientist would knowingly design a study with power in the left shelf. Therefore, all positions described next consider only the middle and right portions of the curve.

The first position corresponds to insuring that a study has power at least in the middle of the curve, say .50 to .70. In some scientific circles, the power value of .80 (β = .20) has achieved nearly the same popularity as α = .05. A power of .50 corresponds to being equally likely to find the effect as to not find the effect. Similarly a power of .75 corresponds to finding the effect three out of four times, and missing the effect one out of four times. Unfortunately the power of a study in the slope region is very sensitive to any small change in effect size, error variance, or sample size. Two or three small differences which all go against the scientist may severely reduce the power.

The second position may be derived from characteristics of the power curve deduced using asymptotic properties of the noncentral distribution function.[2] For the *t* test example, the boundary between the slope and the right shelf corresponds approximately to a study with power greater than .84. This also applies to an *F* test (which includes the *t* test as a special case). Studies with power in the region above the boundary (the right shelf) are not very sensitive to any small change in effect size, error variance, or sample size. Therefore, a desire for robustness of power leads to choosing a study with power of at least .84.

The third position emanates from the wish to control Type I and Type II error rate equally well. This typically leads to requiring power of at least .90 and often .95 or more. Doing

[2] The reader who wishes to examine the details should note that a) noncentral *F* tends to behave like a noncentral χ^2 as denominator *df* increase (17), b) an appropriately scaled noncentral χ^2 tends to behave like a Gaussian random variable as either the noncentrality or *df* parameter increases by itself (17), and c) the shelf boundaries are defined by the points of inflection of the derivative of the power curve (by examination of third derivatives).

so insures that any modest change in effect size, error variance, or sample size, or even a combination of such changes, cannot adversely affect power. Studies of this kind have power values away from the boundary between the slope and the right shelf.

Without consideration of the other issues involved in choosing a power level, we prefer power values in the right shelf. The increase in robustness from moving away from the boundary seems worthwhile. We recommend designing studies with power of approximately .90. Conclusions concerning lack of effect which are based on retrospective power analysis would apparently need to meet the same standard. In this application, the Type II error rate plays the role usually assigned to the Type II error rate. Therefore, an even higher power requirement also seems defensible in order to make a strong claim.

METHODS FOR COMPUTING POWER

Printed Sources

Books provide the most widely available access to power analysis. Probably the most widely used reference in power analysis is by Cohen (4). The book is aimed directly at the practicing scientist, and therefore is centered on explanation with minimal formulas. The 1988 edition is simply a reprinting of the 1977 book. Consequently it lacks coverage of some newer methods. In addition, the tables and methods are based on approximations which lose accuracy as sample sizes get smaller. A more limited handbook (23) has many attractive features. Another treatment of the same topic, although in a rather idiosyncratic manner, can be found in Kraemer and Thiemann (20). All three sources do provide a detailed introduction to the principles and practice of power analysis for a range of practical situations. These books also include useful references if the reader wishes to delve further into a particular method. In summary, using any one of these books is often sufficient and infinitely better than no power analysis. Despite this, lack of coverage of a particular topic should not be taken as a guarantee that no help is available.

Power calculation methods for many techniques are scattered throughout the statistical literature. Unfortunately most authors of available statistics texts failed to include coverage of power methods. The aforementioned books cover the great majority of analysis techniques used by readers of this journal. Certain articles provide additional information that may be useful. Useful tables for the power of the multiple correlation test for the case in which the predictors follow a Gaussian distribution are available in Gatsonis and Sampson's paper (8). The accuracy of popular sample size formulas has been reviewed by Kupper and Hafner (21). An excellent tutorial on power calculations in univariate linear models can be found in O'Brien and Lohr's work (31).

Power calculation methods have been developed only relatively recently for some methods. For example, methods for approximating power for repeated measures ANOVA with multivariate and corrected univariate approaches respectively have been reported (26,27,28).

Computer Software

Producing graphs and tables of power calculations can clearly be done best with computers. A substantial amount of power analysis software has been published. The utility of the programs depends on the computer hardware, the operating system, the presence of other software, and the computing and statistical sophistication of the data analyst. As of this writing, the statistical package corporations have only begun to recognize the need for power analysis integrated with data analysis. The reader should consult with their statistical package's representatives to discover what is available. Such corporations respond positively to user requests when the volume becomes sufficient.

Given the short time during which such information would be accurate, only a shallow survey is provided. It is likely that the most common current computer environment for readers is an IBM compatible microcomputer running PC-DOS. Goldstein reviewed power software for that particular operating system (11). Since that review both SYSTAT and BMDP have released modules. The second most common environment for readers is an Apple Macintosh using the Apple supplied operating system. The SYSTAT module is supported for the Macintosh. Version 2 of JMP, from SAS Institute, has a useful amount of power analysis available.

The widespread use of repeated measures analysis does not correspond to widespread availability of appropriate power software. SPSS MANOVA has some abilities embedded in it. For those able and willing to conduct power analysis in terms of matrix algebra notation, software is distributed with the SAS IML example library to all registered sites. As of this writing, the distribution is expected to begin in Fall, 1992.

CONCLUSIONS

1. Prospective use of power analysis provides substantial advantages. By embedding power evaluation in the habits of design, the scientist minimizes research effort and maximizes research sensitivity. Power analysis stimulates the iterative refinement of scientific hypothesis, design, and analysis.

2. A sensitivity analysis is a necessary component of any power analysis. The sensitivity analysis should account for uncertainty in estimates of parameters, such as the error variance in a t test. Graphs and tables improve not only the clarity of presentation but also the likelihood of an optimal decision about choices of design, analysis, and sample size.

3. Align power calculations with the study. The effort in conducting a power analysis should reflect the money to be spent, the time to complete study, and ethical issues. Power should be calculated for the actual analysis and hypothesis of most interest. When this is not possible, the nearest approximation available should be used.

4. Retrospective use of power analysis provides substantial advantages. Claims of no difference between treatments should include appropriate power analysis. It also can be used to optimize design in planning replications.

5. The choice of effect judged to be of scientific importance should be based on practical relevance and comparison to natural sources of effect.

6. When consideration of ethical and opportunity costs does not preclude it, power should be at least .84, and preferably greater than .90. This corresponds to the right shelf of the power curve.

7. Convenient tables and software are available, at least for the most common applications. Lobbying software vendors may help stimulate better availability. More complex power analyses may require the collaboration of a statistician.

ACKNOWLEDGEMENTS

Supported in part by NICHD Mental Retardation Research Center grant P30-HD03110-22, NCI program project grant P01 CA47 982-04, and NIH Clinical Research Center grant RR00046.

REFERENCES

1. Benignus, V. A.; Kafer, E. R.; Muller, K. E.; Case, M. W. Absence of symptoms with carboxyhemoglobin levels of 16–23%. Neurotox. Teratol. 9:345–348; 1987.
2. Benignus, V. A.; Muller, K. E.; Malott, C. M. Dose-effects functions for carboxyhemoglobin and behavior. Neurotox. Teratol. 12:111–118; 1990.
3. Benignus, V. A.; Muller, K. E.; Smith, M. V.; Pieper, K. S.; Prah, J. D. Compensatory tracking in humans with elevated carboxyhemoglobin. Neurotox. Teratol. 12:105–110; 1990.
4. Cohen, J. Statistical power analysis for the behavioral sciences. (1st ed., reprinting) New York: Academic Press; 1977, 1988.
5. Cooper, H. M. Integrating research: A guide for literature reviews. Newbury Park, CA: Sage, 2nd ed.; 1989.
6. Daniel, W. W. Biostatistics: A foundation for analysis in the health sciences. New York: Wiley, 5th ed.; 1991:195.
7. Edwards, A. L. Statistical methods. New York: Holt, Rinehart, and Winston, 2nd ed.; 1967:236.
8. Gatsonis, C.; Sampson, A. R. Multiple correlation: exact power and sample size calculations. Psychological Bull. 106:516–524; 1989.
9. Glass, G. V. Primary, secondary, and meta-analysis of research. Educational Researcher, 5:3–8; 1976.
10. Glass, G. V.; McGraw, B.; Smith, M. L. Meta-analysis in social research. Newbury Park, CA: Sage; 1981.
11. Goldstein, R. Power and sample size via MS/PC–DOS computers. American Statistician, 43:253–260; 1989.
12. Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 4th ed.; 1965:210.
13. Hayes, W. L. Statistics. New York: Holt, Rinehart, and Winston, 1st ed.; 1963:272.
14. Harbin, T. J.; Benignus, V. A.; Muller, K. E.; Barton, C. N. The effects of low-level carbon monoxide exposure upon evoked cortical potentials in young and elderly men. Neurotox. Teratol. 10:93–100; 1988.
15. Hedges, L. V.; Olkin, I. Statistical methods for meta-analysis. Orlando, FL: Academic Press; 1989.
16. Hunter, J. E.; Schmidt, F. L. Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage; 1990.
17. Johnson, N. L.; Kotz, S. Continuous univariate distributions-2. New York: Houghton Mifflin; 1970:135, 193.
18. Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, CA: Wadsworth, 2nd ed.; 1982.
19. Kleinbaum, D. G.; Kupper, L. L.; Muller, K. E. Applied regression analysis and other multivariable methods. Boston: PWS-Kent, 2nd ed.; 1988:31.
20. Kraemer, H. C.; Thiemann, S. How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage; 1987.
21. Kupper, L. L.; Hafner, K. B. How appropriate are popular sample size formulas? The American Statistician, 43:101–105; 1989.
22. Light, R. J.; Pillemer, D. B. Summing up: The science of reviewing research. Cambridge, MA: Harvard University Press; 1984.
23. Lipsey, M. W. Design sensitivity: Statistical power for experimental research. Newbury Park, CA: Sage; 1990.
24. Maxwell, S. E.; Delaney, H. D. Designing experiments and analyzing data. Belmont, CA: Wadsworth; 1990.
25. Morrison, D. F. Multivariate statistical methods. New York: McGraw-Hill, 2nd ed.; 1976.
26. Muller, K. E.; Peterson, B. L. Practical methods for computing power in testing the multivariate general linear hypothesis. Computational Statistics and Data Analysis, 2:143–158; 1984.
27. Muller, K. E.; Barton, C. N. Approximate power for repeated measures ANOVA lacking sphericity. J. American Statistical Association, 84:549–555; 1989.
28. Muller, K. E.; Barton, C. N. Correction to "Approximate power for repeated measures ANOVA lacking sphericity." J. American Statistical Association, 86:255–256; 1991.
29. Muller, K. E.; Barton, C. N.; Benignus, V. A. Recommendations for appropriate statistical practice in toxicology. Neurotox. 5: 113–126; 1984.
30. O'Brien, R. G.; Kaiser, M. K. MANOVA method for analyzing repeated measures designs: An extensive primer. Psych. Bull. 97: 316–333; 1985.
31. O'Brien, R. G.; Lohr, V. I. Power analysis for linear models: the time has come.,Proceedings of the ninth annual SAS users group international conference, Cary, NC: SAS Institute; 1984.
32. Raudenbach, S. W. Review of "Methods of Meta-Analysis: Correcting Error and Bias in Research Findings." J. American Statistical Assoc. 86:242–244; 1991.
33. Rosenthal, R. Meta-analytic procedures for social research. Newbury Park, CA: Sage; 1984.
34. Timm, N. H. Multivariate analysis. Monterey, CA: Brooks/Cole; 1975.