Estimation of Power Using Simulation Techniques

Alan B. Davis, PhD, and Mark B. Mengel, MD, MPH
Oklahoma City, Oklahoma

A statistical test provides a succinct summary of the investigation of a null hypothesis. A particular feature of a statistical test is protection against type I error, the error of rejecting a hypothesis when it is true. The investigator rejects the null hypothesis when the probability of occurrence of the statistic is smaller than a predetermined probability, alpha. A common value used for alpha is .05, which implies that the investigator is allowing a probability of type I error to be .05.

A type II error occurs when a hypothesis is not rejected when it should be. Such an error may occur when two different interventions differ in their outcome, but the difference is so small that it is not statistically significant. The probability of occurrence of this type of error is beta.

Beta error increases as alpha error decreases because reducing alpha increases the difficulty of rejecting the null hypothesis. Other determinants of beta error include the sample size and the effect size. Effect size is the magnitude of effects under investigation. If the investigation is comparing two interventions that have an important difference in outcome, then it is valuable to know that a research study design will result in the rejection of the null hypothesis. The probability of this occurrence is known as the "power" of a statistical test. Power is calculated by evaluating the probability that the statistic exceeds the critical value of alpha for a given sample size and alternative to the null. Just as a value of .05 is a nominal standard for alpha, a nominal standard for power is .8.

A researcher would not want to commit time and effort to a study that has low power, such as .5, since this would imply only a 50/50 chance of showing statistical significance. In a paper on the uses of power calculation, Fagley[1] described a high power as .8 or above. Additionally, Cohen[2] points out that in setting criteria for alpha and beta error, the researcher is placing a relative importance on the types of error that are possible. If $\alpha = .05$ and $\beta = .20$ (ie, power = .8) then it is four times as important to avoid a type I error than a type II error.

Both Fagley and Cohen advocate the importance of doing a power analysis prior to a research study, but also advocate doing a power analysis post hoc to evaluate the importance of a study if nonsignificant results are reported and a formal power analysis was not done.

Procedures for determining power in research designs are available for the most common statistical procedures. Power tables for *t* tests, contingency tables, analysis of variance, and goodness-of-fit tests are available in Cohen.[2] Little help is available, however, to determine the power of multivariate analysis techniques, such as log-linear modeling, logistic regression, and survival analysis.

METHODS

This paper describes a Monte Carlo technique for determining the power of a multivariate statistical test. This technique involves the following:

1. Decide what hypotheses are tested in the investigation and, specifically, what statistical tests will be used.
2. Determine what alternative or rate difference from the null hypothesis (effect size) is desirable to detect.
3. Simulate a number (*r*) of experiments, and analyze them according to number 1 above.

A sampling estimate of power ($\hat{\omega}$) is then obtained, which is the proportion of simulations resulting in rejection of the null hypothesis, with a confidence interval about $\hat{\omega}$ based on a sample of size *r*.

In determining the number of experiments to simulate, it is necessary to choose the desired accuracy for the estimation of power. A 95% confidence interval around an estimated power $\hat{\omega}$ is $\pm 1.96\ (\hat{\omega}[1 - \hat{\omega}]/r)^{1/2}$. Suppose an accuracy (c) of +.05 around $\hat{\omega}$ is desired. Then the confidence interval equation can be solved, setting $\hat{\omega} = .5$ (the quantity $\hat{\omega}[1 - \hat{\omega}]$ is maximum at $\hat{\omega} = .5$ yielding a "worst-case" estimate) as follows:

$$c = 1.96\ (\hat{\omega}[1 - \hat{\omega}]/r)^{1/2}$$

therefore,

Continued on page 581

Submitted, revised, January 9, 1990.

From the Department of Family Medicine, The University of Oklahoma Health Sciences Center, Oklahoma City. Requests for reprints should be addressed to Alan B. Davis, PhD, Department of Family Medicine, The University of Oklahoma Health Sciences Center, PO Box 26901, Oklahoma City, OK 73190.

Continued from page 578

TABLE 1. CURE RATES OBSERVED, BY TREATMENT GROUP

Group			
Dose	Sex Partner Treated	Percent Cured	Number of Patients
7-day course	Yes	94	33
7-day course	No	75	34
Single dose	Yes	91	34
Single dose	No	65	37

$$r = (1.96)^2 \hat{\omega}\,(1 - \hat{\omega})/c^2 = (1.96)^2\,(.5/.05)^2 = 384.16,$$

or a maximum of 385 simulations is required to estimate $\hat{\omega}$ within ±.05.

EXAMPLE

The following example is included to show how this technique can be used to calculate the statistical power of a clinical investigation in which no significant difference in treatment outcome was found. An investigation was performed to examine the effectiveness of single-dose metronidazole therapy in the treatment of patients and their sexual partner with bacterial vaginosis, a superficial vaginal infection caused by a mixture of anaerobic organisms.[3] As the patient received either a single dose or a 7-day course of metronidazole and the partner received either a single dose or placebo therapy, four treatment groups were formed (Table 1). The design of this trial is a two-factor design, with a dichotomous outcome. A log-linear model was used to analyze significance of effect. Partner treatment with a single dose of metronidazole significantly increased the cure rate of women with bacterial vaginosis ($P < .05$), while the effect of patient treatment was nonsignificant. Before concluding that a 2-g single dose was just as effective as a 7-day course of metronidazole in patients with bacterial vaginosis, a power analysis was performed to determine whether the above conclusion was justified.

In performing a power analysis, one must specify a clinically important difference in cure rates. In general, ineffective therapies for bacterial vaginosis, such as tetracycline or triple sulfa vaginal cream, have cure rates of about 50%.[4] Traditionally the power of a study has been assessed using either a 25% or 50% difference in cure rates.[5] The power of the bacterial vaginosis study to find a 25% difference in cure rates, the more stringent criterion, was assessed so that clinicians would be assured that single-dose metronidazole therapy was at least more effective than triple sulfa vaginal cream. The cure rate of

TABLE 2. RATES FOR ALTERNATIVE TO NULL HYPOTHESIS

Group				
Dose	Sex Partner Treated	Percent Cured	How Determined	Odds Ratio
7-day course	Yes	94	Baseline	5.22
7-day course	No	75	Baseline	
Single dose	Yes	70.5	75% of baseline	1.89
Single dose	No	56	75% of baseline	

the 7-day treatment course was selected as the baseline, 94% if the woman's sexual partner was treated, 75% if her sexual partner was not treated. A cure rate of 70.5%, if the woman received single-dose therapy with her sexual partner, and 56%, without sexual partner treatment, were calculated from the baseline cure rates using the 25% difference criteria (Table 2). These calculated significant cure rates were then used in the computer simulation.

Simulating the above clinical trial involved drawing a random number using a SAS[6] pseudo-random number generator uniformly distributed on the interval 0 to 1 for each patient and then determining the response variable (cure). If the random number was smaller than the hypothesized probability of cure for that category of treatment, then the response was coded as a cure. For instance, if cure rate is .75 and the random number drawn was .749, then the response was coded as a cure. The sample sizes for the four treatment groups were chosen to agree with the experiment as it has been reported.[3]

Assuming these percentages to represent the "true" probabilities of success in the four categories, 385 data sets were generated using a random data-generation technique. Each data set was analyzed using the log-linear model to test the significance of patient treatment, partner treatment, and the interaction of patient and partner treatment. The importance of the interaction term is that the odds ratios differ in the subtables for the hypothesized alternatives of a 25% difference in cure between 7-day and single dose. For the 7-day treatment group, the odds ratios of being cured when the partner is treated to being cured when partner is not treated is 5.22, while for single dose, the odds ratio is 1.89.

The log-linear analysis was considered to show significant effect of patient dosage if this factor provided either a significant main effect or interaction in the analysis of variance table.

In 385 simulations, the patient treatment effect was shown to be significant at $\alpha = .05$, 79.7% of the time. A 95% confidence interval estimate of power is therefore (.757, .837).

These data could have been analyzed using logistic regression. By substituting this analysis in the simulation

Continued on page 584

Continued from page 581

program, an estimate of the power of logistic regression was .745, with a confidence interval of (.71, .79). Both types of analysis on the hypothesized alternative show that the power is high in detecting a significant effect for patient treatment in this study.

CONCLUSIONS

The Monte Carlo method allows great flexibility in the simulation of multivariate research data. In this discussion the method was used to test the power of a research study having a specified number of cases in each treatment group. This method could also be used to estimate the power of other multivariate analysis strategies, such as multiple regression, if the researcher was able to estimate multivariate distributions and was careful to honor the assumptions underlying those tests.

A limitation to this method is that there is no direct way to estimate sample size given the other parameters of the design, ie, power, alpha, effect size, and a model to estimate and test effects. One could estimate sample size indirectly by starting with a sample size equivalent to the power of a univariate test that does not adjust for treatment categories or factor variables. A multivariate experimental design should increase the power of the analysis, so the initial estimate will probably exceed the desired power. One could then adjust the sample size or subset of the initial data set until a sample size with the desired power is attained.

References

1. Fagley, NS: Applied statistical power analysis and the interpretation of nonsignificant results by research consumers. J Counsel Psychol 1985; 32(3):391–396
2. Cohen J: Statistical Power Analysis for the Behavioral Sciences. New York, Academic Press, 1977
3. Mengel MB, Berg AO, Weaver CH, et al: The effectiveness of single-dose metronidazole therapy for patients and their partners with bacterial vaginosis. J Fam Pract 1989; 28:163–171
4. Pheifer TA, Forsyth PS, Durfee MA, et al: Nonspecific vaginitis: Role of *Haemophilus vaginalis* and treatment with metronidazole. N Engl J Med 1978; 298:1429–1434
5. Freiman JA, Chalmers TC, Smith H, Kuebler RR: The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. N Engl J Med 1978; 299:690–694
6. SAS User's Guide: Basics, Version 5 Edition. Cary, NC, SAS Institute, 1985, pp 279