TECHNICAL NOTES

Sample Sizes and Power Computation for Clinical Intervention Trials

Nancy Kline Leidy
Lisa A. Weissfeld

To date, a large portion of nursing's empirical efforts has been directed toward understanding clinical phenomena through the identification and explication of theoretical concepts and relationships. These endeavors have traditionally, and quite appropriately, employed descriptive, cross-sectional research designs in an attempt to quantify phenomena, explore concept independence and covariation, and develop or test causal models. More important, these investigations have provided the groundwork for the next phase in the evolution of nursing science: The widespread use of experimental design to test the efficacy of theoretically and empirically based nursing interventions. Experimental research, in the form of clinical trials or intervention studies with clearly identified nursing protocols and outcome criteria, is essential to the advancement of nursing science and practice. The purpose of this article is to discuss two crucial elements in planning intervention studies and interpreting their results: the estimation of sample size and the computation of achieved statistical power (i.e., power analysis).

Pocock (1983) proposed five key questions that must be addressed when determining sample size for an experimental study:

1. What is the main purpose of the study?
2. What is the primary client outcome measure?
3. How will the data be analyzed to detect an intervention effect?
4. What type of results does one anticipate with standard intervention?
5. How small an intervention difference, or effect size, is considered to be of sufficient clinical importance to detect, and with what degree of confidence?

Each question is designed to guide the investigator toward the ultimate decision: the specification of a sample size which will ensure a reasonable chance of detecting a

Nancy Kline Leidy, R.N., Ph.D., Les Balkans, Givrins, Switzerland; *Lisa A. Weissfeld*, Ph.D., Department of Biostatistics and the Center for Nursing Research, University of Michigan, Ann Arbor.

realistic and theoretically or clinically relevant intervention difference, if one exists. Determination of sample size is not only a scientific concern but an ethical imperative. As Bird and Hall (1986) pointed out, studies that do not include enough observations to provide a reasonable chance of detecting an effect of substantive importance are virtually useless to the scientific community, a waste of client time, and a dissipation of societal resources.

POWER ANALYSIS: A BRIEF REVIEW

The main goal of a clinical trial is to test the efficacy of carefully defined nursing interventions by comparing two or more groups with respect to an observable endpoint, or outcome variable. (For purposes of this article, a single outcome variable and random assignment to groups are assumed.) This goal is achieved by a priori identification of a "reasonable" effect size (i.e., the minimal difference between groups which is considered to be clinically meaningful) and a statistical test of hypothesis. A statistical test is subject to two different types of error. Type I error, with probability α, occurs when the null hypothesis of no difference between groups, or no intervention effect, is falsely rejected. By convention, the value of α is usually taken to be .05. Type II error, with probability β, involves a situation in which the hypothesis of no difference between groups is falsely accepted (i.e., a conclusion of no intervention effect is reached when, in fact, an effect exists). The probability that the test will reject the null hypothesis correctly is $1 - \beta$, the power of the test. Clearly, for the specified level of α, it is in the best interest of the investigator to maximize power, thus increasing the probability that the hypothesis of no clinically meaningful nursing intervention effect (H_0), will be appropriately rejected. Sample size and power computation for a clinical intervention trial are dependent on effect size and the probability of Type I error which is specified by the investigator.

The nature of the outcome variable, or endpoint, in a clinical trial is also an important component of sample size and power computation. Intervention outcome variables can be categorized as continuous or dichotomous measures. A continuous endpoint is one in which the outcome variable is measured at an interval level and assumed to be normally distributed, while a dichotomous outcome occurs in a situation in which the endpoint can be classified into one of two qualitative categories, commonly success or failure.

SAMPLE SIZE COMPUTATION FOR TRIALS WITH A CONTINUOUS OUTCOME MEASURE

Consider an intervention study where two groups are being compared with an outcome measure that is continuous. Let $\bar{X}_1$ be the mean value of this measure for the control group and $\bar{X}_2$ be the value of this measure for the experimental group. Assume

that $\bar{X}_1$ is normally distributed with mean μ_1 and variance σ_1^2, while $\bar{X}_2$ is normally distributed with mean μ_2 and varianceσ_2^2. The hypothesis to be tested is formulated as:

$$H_0: |\mu_1 - \mu_2| \leq \varepsilon$$

$$H_1: |\mu_1 - \mu_2| > \varepsilon,$$

where $|\mu_1 - \mu_2|$ is the effect size and ε is the minimum clinically relevant effect size of interest. The *t* test is typically used to test this hypothesis, where the *t* statistic employs the pooled estimate of the common variance, S_p^2 with $N - 2$ degrees of freedom.

To estimate sample size and power, let Q_1 be the proportion of the sample to be assigned to the control group and Q_2 be the proportion of the sample to be assigned to the experimental group. If N is the total sample size required for the study, then the number of clients to be randomized to the control group, n_1, is NQ_1. Similarly, the number of clients to be randomized to the experimental group, n_2, is NQ_2, where $Q_1 + Q_2 = 1$. This allows for the existence of unequal group sizes, should limited or restricted sampling resources and/or ethical considerations make this a necessity. Let $Z\alpha$ denote the standard normal deviate at the designated significance level, α, and let Z_β denote the standard normal deviate at level β (obtained by referring Z_β to a normal distribution table) so that the power $(1 - \beta)$ is determined by subtracting the associated probability level β from 1. For example, $Z_\beta = .84$ for a β level of .20 and a power of .80.

As above, let ε denote the effect size, or the intervention – control difference $(\mu_2 - \mu_1)$ for the endpoint of interest. An estimate of variability in the outcome variable is also needed. This pooled variance estimate, denoted by S_p^2, can be obtained from previous studies reported in the literature or through a pilot study. It is advisable to use the largest S_p^2 expected so that the power and sample size estimates are conservative enough to ensure a reasonable outcome (Lachin, 1981). In situations in which the power of a previously conducted trial is being evaluated, the observed sample and effect sizes and the obtained estimate of the variance are employed in the calculation.

Under these assumptions, the total sample size N required to ensure adequate power to detect a clinically relevant and significant difference between groups is estimated by (Cochran, 1983):

$$N = \frac{S_p^2 (Q_1^{-1} + Q_2^{-1}) (Z_\alpha + Z_\beta)^2}{\varepsilon^2} \tag{1}$$

The power of a study to detect a relevant difference for a specified sample size is given by solving equation (1) in terms of Z_α (Cochran, 1983):

$$Z_\beta = \frac{|\varepsilon| \sqrt{N} - Z_\alpha \sqrt{(Q_1^{-1} + Q_2^{-1})}}{S_p \sqrt{(Q_1^{-1} + Q_2^{-1})}}. \tag{2}$$

From Equation 1 it can be seen that as ε increases, the sample size, *N*, required to obtain statistically significant results decreases. Similarly, according to Equation 2, as ε and/or N increases, power will be enhanced. That is, Z_β will increase, the associated probability, β, will decrease and, clearly, 1 – β, or power, will rise. Finally, the required *N* is reduced and power maximized when the two groups have approximately equal sample sizes (i.e., Q_1 approaches Q_2; note $Q1^{-1} + Q_2^{-1} = 4.0$ when $Q_1 = Q_2$).

The same sample size and power estimation procedures can be employed when subjects in the two groups serve as their own control, with measures before and after treatment. Under these circumstances, the individual observations consist of the pretest-posttest differences and the estimate of the variance of the differences (S_d^2) is employed in calculating the *t* statistic. Thus, in Equations 1 and 2, $\varepsilon = |\delta_1 - \delta_2|$, where $\delta_1 = (\mu_{1B} - \mu_{1A})$ and $\delta_2 = (\mu_{2B} - \mu_{2A})$, and Sα is substituted for S_p. Other adjustments to these computations are needed if a nonparametric test, such as the Wilcoxon rank sum test, is used to test the null hypothesis of no treatment difference (Cohen, 1977; Lehmann, 1975).

A frequently identified problem with sample size computation for continuous outcome measures is that the effect size, ε, and variance, S^2, cannot be estimated. Under these circumstances, the investigator can redefine the problem from one of a continuous mean response difference ($\mu_1 - \mu_2$) to a dichotomous outcome measure (the probability of a successful response in the control group versus the intervention group) in order to estimate sample size (Pocock, 1983).

SAMPLE SIZE COMPUTATION FOR TRIALS WITH A DICHOTOMOUS OUTCOME MEASURE

For an intervention study where the outcome variable is dichotomous, sample size is determined by specifying the expected proportion of successful outcomes in the control group π_1, the expected proportion of successful outcomes in the experimental group, π_2, the minimal relevant difference between these proportions ($\varepsilon = \pi_1 - \pi_2$), the significance level α, and the power (i.e., Z_α). Alternatively, power is determined by ε, α, and *N*. The hypothesis is formulated as

$$H_0: \pi_1 \leq \pi_2 + \varepsilon \text{ or } |\pi_1 - \pi_2| \leq \varepsilon$$

$$H_1: \pi_1 > \pi_2 + \varepsilon \text{ or } |\pi_1 - \pi_2| > \varepsilon$$

and is tested using the standard test for the difference between two proportions, most commonly, the χ^2 test.

The following equations are used to determine sample size and power in studies with a dichotomous outcome measure and equal group sizes (Lachin, 1981):

$$N = \frac{(Z_\alpha - Z_\beta)^2 \pi^*(1 - \pi^*)}{(\pi_2 - \pi_1)^2} \tag{3}$$

$$Z_\beta = \frac{\sqrt{N} \mid \pi_2 - \pi_1 \mid}{2\sqrt{\pi^*(1 - \pi^*)}} - Z_\alpha, \tag{4}$$

where $\pi^* = Q_1\pi_1 + Q_2\pi_2$. Note that Equations 3 and 4 depend on specifying the values of π_1 and π_2 rather than simply estimating the effect size, ε. Equation 3 also shows that as the success rate in the control group, π_1, decreases, the required sample size is reduced and the power enhanced as long as $\pi_1 < \pi_2$. Because misspecification of the proportions π_1 and π_2 may result in an underestimation of sample size and an overestimation of power, the largest realistic value possible for π_2 should be specified, under the constraints that $\pi_1 > \pi_2$ and $\pi_2 < 0.5$ (Lachin, 1981).

As an example, suppose π_1 (the proportion of success in the control group) = .45 and π_2 (the proportion of success in the experimental group) = .75. Assuming equal sample sizes are possible, $\varepsilon = .30$, $\pi = .30$ and $4\pi(1-\pi) = .84$. Using $\alpha = .05$ (one-sided) and $\beta = .10$ (90% power), the optimal sample size would be 80, with 40 subjects randomly assigned to the two groups. Should the study be conducted with only 30 subjects, 15 in each group, the power of the study to detect $\varepsilon = .30$ with $\pi_1 = .45$ is only 56% ($Z_\beta = .148$).

To estimate N or Z_β when the two groups are not equal, that is, $Q_1 \neq Q_2$, the following equation is used (Lachin, 1981):

$$\sqrt{N} \mid \pi_2 - \pi_1 \mid = Z_\alpha \sqrt{\pi^*(1 - \pi^*)\,(Q_1^{-1} + Q_2^{-1})} + Z_\alpha \sqrt{\pi_2 (1 - \pi_2)\, Q_2^{-1} + \pi_1(1 - \pi_1)\, Q_1^{-1}} \tag{5}$$

The investigator solves for either N or Z_β.

ADDITIONAL CONSIDERATIONS

Dropout Rates and Noncompliance

A primary consideration in the determination of sample size is the projected rate of dropout (R), that is, the number of subjects who terminate participation or do not adhere to the protocol. Lachin (1981) suggested a simple adjustment, $N_a = N(1\text{-}R)^2$, where N_a is the sample size required with dropouts and N is the sample size calculated without dropouts. An alternative method focusing on the effect of nonadherence to protocol on sample size estimates is proposed by Schork and Remington (1967). This method relies on a table to provide the adjustment for nonadherence. Clearly, these methods do not negate the need for careful attention given to the planning and

administration of the study to minimize additional factors that might alter the desired effect (e.g., noncompliance or lack of control of influential factors).

Parameter Combinations to Determine Sample Size

Sample size determination during the proposal phase may lead to estimates that are larger than originally anticipated. In this situation, it may be advisable to consider a higher α or β risk (or both), or a larger effect size, depending on the nature of the intervention and clinical outcome measure. Alternatively, if it is clear that acceptable definitive results cannot be obtained with the available resources, the investigator should consider addressing a more realistic, alternative research question (Frieman, Chalmers, Smith, & Kuebler, 1978).

Number of Intervention Groups

Because the power of a study is dependent on the number of subjects per group, rather than the total number of subjects in the trial (see Equations 1, 2, and 5), investigators anticipating difficulty in subject recruitment may employ two-group designs (Pocock, 1983). Although the temptation may be great to propose the testing of multiple therapeutic interventions in a single study, failure to achieve the required sample size will yield results which are indeterminate and thus of less scientific value than more definitive findings that would have been achieved by comparing two carefully identified nursing interventions. Clearly, two-group intervention studies, with replication and refinement, play an important role in understanding the efficacy of nursing interventions, thereby contributing more to the accumulation and advancement of nursing knowledge than multigroup studies with low statistical power.

Reporting Results

In general, clinical trials in medicine and nursing have failed to address a priori sample size and power estimation in the published reports of their findings. In a study of 71 negative ($p > .05$) clinical trials published in medical journals, Frieman et al. (1978) found only one paper stating that significance level and power were considered before the start of the trial. Only 18 papers recognized a potentially meaningful trend and 14 suggested a larger sample size. Further, most of the studies reviewed suggested that failure to achieve statistical significance was indicative of no clinically meaningful difference between groups rather than a consequence of insufficient power. In nursing, where clinical trials have been less common, Jacobsen and Meininger (1986) found that only 2 of 42 reports published in three major research journals between 1980 and 1984 estimated the required sample size prior to the study. The authors did not describe the extent to which the investigators interpreted their findings within the

context of statistical power. Clearly, power estimation would be a useful addition to published reports of nursing intervention trials.

SUMMARY

The purpose of this article was to describe basic concepts of sample size and power estimation for planning nursing intervention trials and interpreting their results. Simple mathematical calculations, using the formulas presented here, can be used to estimate the number of subjects required to conduct a study with a designated effect size and level of power. These methods are of great importance, since most funding agencies require sample size and power estimations before a grant is awarded. In general, studies with power lower than .7 or .8 need careful consideration before they are implemented. In these situations, it may be wise to consider various alternatives for obtaining study subjects or deleting treatment groups for investigations involving more than two groups. The formulas presented here can also be useful in estimating the power of published research findings. Through a quick calculation, the consumer of nursing research can critically evaluate the meaning of a negative trial and draw appropriate conclusions for future research and practice.

REFERENCES

Bird, K., & Hall, W. (1986). Statistical power in psychiatric research. *Australian and New Zealand Journal of Psychiatry, 20*, 189-200.

Cochran, W. G. (1983). *Planning and analysis of observational studies.* New York: Wiley.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Frieman, J., Chalmers, T., Smith, H., & Kuebler, R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *New England Journal of Medicine, 299*, 690-694.

Jacobsen, B., & Meininger, J. (1986). Randomized experiments in nursing: The quality of reporting. *Nursing Research, 35*, 379-382.

Lachin, J. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials, 2*, 93-113.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks.* San Francisco: Holden-Day.

Pocock, S. (1983). *Clinical trials: A practical approach.* New York: Wiley.

Schork, M. A., & Remington, R. (1967). The determination of sample size in treatment-control comparisons for chronic disease studies in which dropout is a problem. *Journal of Chronic Disease, 20*, 233-239.