Sample Size

Chirayath M. Suchindran
University of North Carolina, Chapel Hill, North Carolina, USA

Glossary

censoring When individuals who are followed for a fixed duration of time do not experience the outcome of interest by the time the observation period ends. Such observations are called censored observations. For a right-censored observation, all that is known is that the time to the event occurrence exceeds the period of observation.

effect size The difference detected in the end point in a study; depending on the end point, the effect size may be means, regression coefficients, odds ratios, or hazards ratios.

intraclass correlation Used as a measure of homogeneity among elements of a cluster; can be viewed as the correlation among pairs of observations within a cluster.

Type I error An error that occurs when the experimental situation declares that the specified difference is real, when, in fact, this is not true. The probability of a Type I error is known as the level of significance.

Type II error In experimental studies, failure to detect the specified difference (the second kind of error). The power of a statistical test is then the conditional probability that the null hypothesis is correctly rejected when it is false (complement of the second kind of error).

A well-designed scientific study must determine, at the outset, the sample size; the sample must be large enough to provide an adequate number of observations such that the quantities of interest can be estimated with sufficient precision and that any difference of importance is likely to be detected. These determinations are based on sound statistical principles. The methods for determining the sample size depend on the goals of the study, the types of outcome measures, the planned mechanism of data gathering, and the tolerance in certain error levels. For example, the planned study may be observational or experimental. The planned data gathering may be through a simple random sample of individuals or through other complex sample design. Often, information is collected through complex sample surveys that involve stratification and several stages of clustering, and the quantities of interest may involve ratio and regression estimates. When the sampling scheme involves several levels, the sample size depends on the magnitude of variations at all levels. Intervention studies may involve many baseline measures before intervention starts and several postintervention measurements to determine the effect of intervention. In follow-up studies, it may also be important to adjust the sampling size for missing data, dropouts, and censoring.

Basic Principles

Sampling techniques are used either to estimate statistical quantities with desired precision or to test statistical hypotheses. The first step in the determination of the sample size is to specify the design of the study (simple random samples of the population, stratified samples, cluster sampling, longitudinal measurement, etc.). If the goal is statistical estimation, the endpoint to be estimated and the desired precision would be specified. The desired precision can be stated in terms of standard error or a specified confidence interval. If the goal is to conduct statistical testing, the determination of sample size will involve specifying (1) the statistical test employed in testing the differences in end point, (2) the difference in the end point to be detected, (3) the anticipated level of variability in the end point (either from previous studies or from theoretical models), and (4) the desired error levels (Type I and Type II errors). The value of increased information in the sample is taken into consideration in the context of the cost of obtaining it. Guidelines are often needed for specifications of effect size and associated variability. One strategy is to take into account as much available prior information as possible. Alternatively, a sample size is selected in advance and the information (say, power or effect size) that is likely to be obtained with that sample size is examined. Large-scale surveys often aim to gather many items of information. If a desired degree of precision is prescribed for each item, calculations may lead to a number of different estimates for the sample size. These are usually compromised within the cost constraint. Sample size determinations under several sampling designs or experimental situations are presented in the following sections.

Simple Random Sampling

A simple random sample (SRS) is the simplest form of probability sample. As stated earlier, the goal of the study may be to estimate a quantity with a desired precision (defined as the variance or the deviance from the population mean) or to test a hypothesis about the mean. Each of the situations can be formally examined under the SRS scheme as follows. Assume that there is population of finite size N from which it is desired to draw a sample of size n. In the first scenario, the goal is to estimate the mean of a quantity with a desired variance V^2. An appropriate value of n can be determined by examining the theoretical value of the variance of the sample mean with the desired variance. From sampling theory, it is known that the sample mean $\bar{y}$, under simple random sampling without replacement has a variance $[(1 - n/N)/n]S^2$, where S^2 is the element variance in the population. Equating the desired variance V^2 to the theoretical value, the desired sample size can be obtained as $n = n'/(1 + n'/N)$, where $n' = S^2/V^2$. If the finite population correction can be ignored, the sample size will be exactly the ratio of the element variance to the desired variance. In another scenario, the precision is expressed differently in terms of margin of errors. The margin of error specification states the closeness of the sample mean to the population mean. Let μ denote the population mean; the aim is to estimate μ with a sample mean within a specified deviance. The specification is made as $P(|\bar{y} - \mu| \leq \varepsilon) = 1 - \alpha$. In this specification, ε is the margin of error and α is the level of significance. Using the results on confidence intervals for sample means obtains an equation connecting the margin of error and sample size as follows:

$$\varepsilon = Z_{\alpha/2}\sqrt{1 - n/N}\,\left(S/\sqrt{n}\right), \tag{1}$$

where $Z_{\alpha/2}$ represents the $(1 - \alpha/2)$th percentile of the standard normal distribution. Writing $n' = Z^2_{\alpha/2}(S^2/\varepsilon^2)$, it can be seen from Eq. (1) that, in the case of sampling with replacement, the required sampling size will be n'. When sampling is done without replacement, the solution of Eq. (1) gives the required sampling size as $n = n'/(1 + n'/N)$. In these examples, sample size calculations are specified as an estimation of the mean of the population with a specified error. In a third scenario, it is possible to specify the estimation of sampling size as a formulation of one sample test of a mean for which the null hypothesis is $\mu + \mu_0$ and the alternative hypothesis is $\mu + \mu_1$. With an assumed level of significance of α and power $1 - \beta$, the required sample size can be obtained under the assumption of normality as follows:

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2}{(\mu_1 - \mu_0)^2} S^2. \tag{2}$$

The various formulations of sample size calculations show the need to have some prior knowledge of the effect size $(\mu_1 - \mu_0)$ and variability S^2 in carrying out the calculations. A number of suggestions have been made in the literature as to how to make reasonable guesses of these values. These include searching past studies that include similar variables or conducting pilot studies. Pilot studies that are too small are less dependable than expert guesses. A reasonable guess can be made concerning the variability in terms of the coefficient of variation. The formulas for sample size calculations can then be expressed in terms of coefficients of variation to obtain the required sample sizes. When the quantity to be estimated is a proportion (P), variance based on a binomial model can be utilized. The term P $(1{-}P)$ is not sensitive to changes in the middle range of P $(0.2 - 0.8)$, and generally, a reasonable guess of the value of P can be made.

Stratified Sampling

Stratification (or blocking) of the study population is often performed prior to sampling in order to increase the precision of the estimate of the quantity of interest. A procedure similar to the simple random sampling case requires knowledge of the variability within each stratum. Such information is seldom available. The concept of "design effect" has been introduced to simplify the calculations. The design effect (denoted as deff) is defined as the ratio of the variance of an estimate under a sampling plan to the variance of the same estimate from a simple random sample with same number of observation units. The sampling plan could be a stratified sampling or other complex sample designs. The design effect is a measure of the precision gained or lost by use of the more complex design instead of a simple random sample. If the design effect can be guessed, it is necessary to estimate the sample size using a simple random sample, as shown in the previous section, and multiply this sample size by deff to obtain the sample size needed under the complex design. Thus, in order to estimate the population mean of a continuous variable with margin of error specified and use of stratified sampling, the required sample size n is obtained using a modification of Eq. (1) (ignoring finite population correction):

$$n = Z^2_{\alpha/2}\left(\frac{S^2}{\varepsilon^2}\right) \times \text{deff}. \qquad (3)$$

The value of the design effect can be obtained from previous surveys or through pilot studies. Once the overall sampling size is determined, the allocation of the samples to strata must be considered. Two methods are generally proposed in the literature. In proportional allocation, the sampling units are allocated in proportion to the size of the stratum. When the variances of observations within strata are more or less equal across strata, proportional allocation is the best allocation for increasing precision. When the variances vary greatly across strata, an optimum allocation procedure is suggested. When the costs of sampling in each stratum are the same, the sample allocation in a stratum h is proportional to the product N_hS_h, where N_h is the size of the strata and S_h is the standard deviation of observations within the strata (Neyman allocation). One difficulty with the optimal allocation is that the correct information on the variability within the strata is rarely obtained.

In experimental situations in which the goal is to compare a treatment group with a control group, the allocation of the samples in each group can be simplified. Denote the mean and the variance of the first (treatment group) as μ_1 and σ_1^2 and the mean and the variance of the second group as μ_2 and σ_2^2. Also assume that the allocation sample to each group is made in a way such that $n_2/n_1 = r$. Note that in this case, the allocation to two groups is predetermined. Then the required sample size can be calculated as follows:

$$n_1 = \frac{(\sigma_1^2 + \sigma_2^2/r)(Z_{\alpha/2} + Z_\beta)^2}{(\mu_1 - \mu_2)^2} \quad \text{and} \quad n_2 = rn_1, \qquad (4)$$

where α is the desired level of significance and $1 - \beta$ is the power.

Cluster Sampling

Many surveys often employ cluster sampling, whereby sampling units are clusters of elements. In a one-stage cluster sampling, every element within a sampled cluster is included in the sample. In two-stage cluster sampling, a subsampling is done to select elements from the chosen clusters. Elements within a cluster may be very similar. A measure of similarity (or homogeneity) of elements within the cluster is provided by the intraclass correlation coefficient (ICC), which is defined to be the Pearson correlation coefficient of all pairs of observations within the cluster taken over all clusters. The ICC plays an important role in the calculation of sample size. For example, in a single-stage cluster sampling, when all clusters are of equal size, the design effect can be approximated as $1 + (M - 1) \times \text{ICC}$, where M is the size of the cluster. In this case, the number of clusters to be selected is calculated in two stages. First, determine the sample size as if the sampling is done under simple random sampling. Then multiply that sample size by the design effect. Once again, the ICC must be known to complete the calculations, but ICC is seldom known and has to be estimated through pilot studies or derived from the values obtained in similar surveys.

Many intervention studies use group-randomized design to examine the effect of an intervention. For example, school-based studies are often used in drug-use prevention studies. In these studies, schools are randomized to treatment and control groups. The question then arises as to how many schools must be selected for each group. In these trials, a school is considered as a cluster and the intraclass correlation is used as a measure of dependence among students within the school. If the goal is to test the difference in the means of a continuous outcome at a significance level α with a desired power $1 - \beta$, the number of schools (n) to be allocated for each group will be

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2 2S^2[1 + (M - 1) \times \text{ICC}]}{M\Delta^2}, \qquad (5)$$

where M is the size of the school, Δ is the hypothesized difference in mean of the treatment and control schools, and S^2 is the total element variance (which includes both within- and between-persons components of variance). In most situations, the cluster (school) size will not be equal. In such situations, the size M is replaced by an average of cluster sizes (usually a harmonic mean of the cluster sizes).

Repeated Measures Design

Experimental studies often involve repeated measurements of outcome measures. For example, for comparison of an intervention group with a control group, intervention studies make baseline measurements of outcome before the intervention and then repeat the measurements one or more times after implementation of the intervention. The sample size requirements depend on the type of hypothesis to be tested, the number of pre- and postintervention measurements, and the level of correlations among observations from the same individual. Under this scenario, sample size formulas have been developed for three possible methods of analysis—namely, (1) a simple analysis using the mean of each individual's postintervention measures as the summary measure, (2) a simple analysis of each individual's difference between means of postintervention and preintervention measurements, and (3) using the preintervention measurements as a covariate in a linear model for comparing the intervention comparison of postintervention means.

Repeated measures data are considered as correlated observations and the generalized estimating equation (GEE) method is employed in analyzing such data. Several authors have discussed estimation of sample size when the GEE method is involved as a tool of analysis; one study provides an approach to estimate sample size for two-group repeated measures when the correlation structure among the repeated measures is unknown.

Follow-up Studies

Follow-up studies usually begin with assigning individuals to an intervention or control group; the individuals are then followed for a fixed period or until the event of interest occurs. The objective of this study design is to detect a change in the rate of occurrence of the event (hazard) in the intervention group in relation to that of the control group. In this study situation, it is possible that some individual observations will be censored; this means that some individuals may not experience the outcome of interest by the time the study is terminated. For censored observations, all that is known is that the time to the event exceeds the duration of observation. The desired sampling size is the minimum number of individuals required to detect a specified change in the hazards ratio. A simple formula has been developed to calculate the required sample size. Let P_I and P_C denote the proportion of individuals assigned, respectively, to the intervention and control group. Let the ratio of the hazard function of individuals in the intervention group to that of the control group be a constant denoted by Δ. Then the total number of individuals required for the study can be expressed as follows:

$$n = \frac{1}{d}\,\frac{(Z_\beta + Z_{1-\alpha})^2}{P_I P_C \log^2 \Delta},$$

where d is the proportion of individuals expected to experience the event of interest. As before, $Z_{1-\alpha}$ and Z_β denote $1-\alpha$ and β percentiles of the normal distribution. The determination of d requires some additional information. Let f denote the planned follow-up time. Often, there is some prior information available about the rate of event occurrence in the control group. Suppose that $S_C(f)$ denotes the probability that an individual in the control group does not experience the event by time f. Then the proportion of individuals in control group experiencing the event by the follow-up time f is $d_C = 1 - S_C(f)$. Thus, under the postulated hazards ratio Δ, the proportion of individuals expected to experience the event in the intervention group is $d_I = 1 - (1 - d_C)^{1/\Delta}$. Then $d = P_C d_C + P_I d_I$.

Epidemiologic Study Designs

Epidemiological studies often use study designs that require special formulas for sample size determinations. For example, in a case-control study, a sample of people with an end point of interest (cases) is compared to a sample without and end point of interest (controls). In this case, the sampling is performed with stratification according to the end point, which is different from the usual stratified sampling. A case-control design will lead to the calculation of an odds ratio as an approximate relative risk, and the sample sizes are determined using an odds ratio as the index of interest. To prevent confounding effects, matched case-control studies (in which the cases and controls are matched at a level of a potentially confounding factor) are sometimes used. Sample size calculations for such designs have been described. Other epidemiologic designs that require special formulas for sample size calculations include nested case-control studies and case-cohort designs.

Covariate Adjustments

Frequently, studies will have end points with regression adjustments for one or more covariates. Many of the sample size calculation formulas can be easily modified to take this situation into account. For example, consider a simple logistic regression situation for which the goal is to examine the relationships of a covariate with a binary response. The model setup is $\log[p/(1-p)] = \beta_0 + \beta_1 x$, where x is a covariate. The sample size determination is made to test the null hypothesis $\beta_1 = 0$. When x is a continuous covariate, the required sample size can be obtained as follows:

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2}{P^*(1-P^*)\beta^{*2}},$$

where P^* is the event rate at the mean of the covariate x and β^* is the effect size to be tested. Simple modifications are needed when the covariate is binary or when additional covariates are included.

Conclusion

Sample size determination is an integral part of any well-designed scientific study. The procedure to determine sample size depends on the proposed design characteristics including the nature of the outcome of interest in the study. There exists a vast amount of literature on the topic, including several books. The modern computer environment also facilitates determination of sample size; software designed exclusively for this purpose is available. Many of the procedures depend on the normality assumption of the statistic. Modern computer-intensive statistical methods give some alternative procedures that do not depend on the normality assumption. For example, many people working in this field of study now prefer to use bootstrap procedures to derive the sample size. Uncertainty in specifying prior information of effect size has led to Bayesian approaches to sample size determination. These computer-intensive procedures seem to have several advantages over many of the conventional procedures of estimating sampling size.

See Also the Following Articles

Age, Period, and Cohort Effects • Clustering • Population vs. Sample • Randomization • Stratified Sampling Types • Type I and Type II Error

Further Reading

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *J. Am. Statist. Assoc.* **91,** 14–28.

Cohen, J. (1988). *Statistical Power for Behavioral Sciences.* Lawrence Erlbaum Assoc., Mahwah, New Jersey.

Ejigou, A. (1996). Power and sample size for matched case-control studies. *Biometrics* **52,** 925–933.

Elashoff, J. D. (2000). *NQuery Advisor. Release 5.0.* Statistical Solutions, Ltd., Cork, Ireland.

Frison, L., and Pocock, S. (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implication for design. *Statist. Med.* **11,** 1685–1704.

Hsieh, F. Y., Bloch, D. A., and Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statist. Med.* **17,** 1623–1634.

Joseph, L., Burger, R. D., and Belisle, P. (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statist. Med.* **16,** 769–789.

Jung, S., and Ahn, C. (2003). Sample size estimation for GEE method for comparing slopes in repeated measurement data. *Statist. Med.* **22,** 1305–1315.

Kish, L. (1995). *Survey Sampling.* John Wiley, New York.

Lenth, R. V. (2001). Some practical guidelines for effective sample size calculations. *Am. Statistic.* **55,** 187–193.

Lenth, R. V. (2003). *Java Applets for Power and Sample Size.* Available on the Internet at www.cs.uiowa.edu

Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized models. *Biometrika* **73,** 13–22.

Liu, G., and Liang, K. Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics* **53,** 937–947.

NCSS. (2002). *Power Analysis and Sample Size Software*. Available on the Internet at www.ncss.com

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73,** 1–11.

Rochon, J. (1991). Sample size calculations for two-group repeated measures experiments. *Biometrics* **47,** 1383–1398.

Schoenfeld, D. A. (1983). Sample size formula for the proportional hazards regression model. *Biometrics* **39,** 499–503.

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

Thomas, L., and Krebs, C. J. (1997). A review of statistical power analysis software. *Bull. Ecol. Soc. Am.* **78**(2), 126–139.

Troendle, J. F., and Yu, K. F. (2003). Estimation of sample size for reference interval studies. *Biometr. J.* **45**(5), 561–572.

Woodward, M. (1992). Formulae for sample size, power and minimum detectable relative risk in medical studies. *Statistician* **41,** 185–196.

Yafune, A., and Ishiguro, M. (1999). Bootstrap approach for constructing confidence intervals for population pharmacokinetic parameters I: Use of bootstrap standard error. *Statist. Med.* **18,** 581–599.

Zou, K. H., Resnic, F. S., Gogate, A. S., Ondategui-Parra, S., and Ohno-Machado, L. (2003). Efficient Bayesian sample size calculation for designing a clinical trial with multi-cluster outcome data. *Biometr. J.* **45**(7), 825–836.