Radiology



John Eng, MD

**Index terms:**
Radiology and radiologists, research
Statistical analysis

**Published online**
10.1148/radiol.2272012051
**Radiology 2003;** 227:309–313

[1] From the Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University, 600 N Wolfe St, Central Radiology Viewing Area, Rm 117, Baltimore, MD 21287. Received December 17, 2001; revision requested January 29, 2002; revision received March 7; accepted March 13. **Address correspondence to** the author (e-mail: *jeng@jhmi.edu*).

© RSNA, 2003

# Sample Size Estimation: How Many Individuals Should Be Studied?[1]

The number of individuals to include in a research study, the sample size of the study, is an important consideration in the design of many clinical studies. This article reviews the basic factors that determine an appropriate sample size and provides methods for its calculation in some simple, yet common, cases. Sample size is closely tied to statistical power, which is the ability of a study to enable detection of a statistically significant difference when there truly is one. A trade-off exists between a feasible sample size and adequate statistical power. Strategies for reducing the necessary sample size while maintaining a reasonable power will also be discussed.

© RSNA, 2003

How many individuals will I need to study? This question is commonly asked by the clinical investigator and exposes one of many issues that are best settled before actually carrying out a study. Consultation with a statistician is worthwhile in addressing many issues of study design, but a statistician is not always readily available. Fortunately, many studies in radiology have simple designs for which determination of an appropriate *sample size*—the number of individuals that should be included for study—is relatively straightforward.

Superficial discussions of sample size determination are included in typical introductory biostatistics texts (1–3). The goal of this article is to augment these introductory discussions with additional practical material. First, the need for considering sample size will be reviewed. Second, the study design parameters affecting sample size will be identified. Third, formulae for calculating appropriate sample sizes for some common study designs will be defined. Finally, some advice will be offered on what to do if the calculated sample size is impracticably large. To assist the reader in performing the calculations described in this article and to encourage experimentation with them, a World Wide Web page has been developed that closely parallels the equations presented in this article. This page can be found at *www.rad.jhmi.edu/jeng/javarad/samplesize/*.

Even if a statistician is readily available, the investigator may find that a working knowledge of the factors affecting sample size will result in more fruitful communication with the statistician and in better research design. A working knowledge of these factors is also required to use one of the numerous Web pages (4–6) and computer programs (7–9) that have been developed for calculating appropriate sample sizes. It should be noted that Web pages for calculating sample size are typically limited for use in situations involving the well-known *parametric statistics*, which are those involving the calculation of summary means, proportions, or other parameters of an assumed underlying statistical distribution such as the normal, Student *t*, or binomial distributions. The calculation of sample size for nonparametric statistics such as the Wilcoxon rank sum test is performed by some computer programs (7,9).

## IMPORTANCE OF SAMPLE SIZE

In a comparative research study, the means or proportions of some characteristic in two or more comparison groups are measured. A statistical test is then applied to determine whether or not there is a significant difference between the means or proportions observed in the comparison groups. We will first consider the comparative type of study.

Sample size is important primarily because of its effect on statistical *power*. Statistical power is the probability that a statistical test will indicate a significant difference when there truly is one. Statistical power is analogous to the sensitivity of a diagnostic test (10), and one could mentally substitute the word "sensitivity" for the word "power" during statistical discussions.

In a study comparing two groups of individuals, the power (sensitivity) of a statistical test must be sufficient to enable detection of a statistically significant difference between the two groups if a difference is truly present. This issue becomes important if the study results were to demonstrate no statistically significant difference. If such a negative result were to occur, there would be two possible interpretations. The first interpretation is that the results of the statistical test are correct and that there truly is no statistically significant difference (a true-negative result). The second interpretation is that the results of the statistical test are erroneous and that there is actually an underlying difference, but the study was not powerful enough (sensitive enough) to find the difference, yielding a false-negative result. In statistical terminology, a false-negative result is known as a *type II error*. An adequate sample size gives a statistical test enough power (sensitivity) so that the first interpretation (that the results are true-negative) is much more plausible than the second interpretation (that a type II error occurred) in the event no statistically significant difference is found in the study.

It is well known that many published clinical research studies possess low statistical power owing to inadequate sample size or other design issues (11,12). One could argue that it is as wasteful and inappropriate to conduct a study with inadequate power as it is to obtain a diagnostic test of insufficient sensitivity to rule out a disease.

## PARAMETERS THAT DETERMINE APPROPRIATE SAMPLE SIZE

An appropriate sample size generally depends on five study design parameters: minimum expected difference (also known as the effect size), estimated measurement variability, desired statistical power, significance criterion, and whether a one- or two-tailed statistical analysis is planned.

### Minimum Expected Difference

This parameter is the smallest measured difference between comparison groups that the investigator would like the study to detect. As the minimum expected difference is made smaller, the sample size needed to detect statistical significance increases. The setting of this parameter is subjective and is based on clinical judgment and experience with the problem being investigated. For example, suppose a study is designed to compare a standard diagnostic procedure of 80% accuracy with a new procedure of unknown but potentially higher accuracy. It would probably be clinically unimportant if the new procedure were only 81% accurate, but suppose the investigator believes that it would be a clinically important improvement if the new procedure were 90% accurate. Therefore, the investigator would choose a minimum expected difference of 10% (0.10). The results of pilot studies or a literature review can also guide the selection of a reasonable minimum difference.

### Estimated Measurement Variability

This parameter is represented by the expected SD in the measurements made within each comparison group. As statistical variability increases, the sample size needed to detect the minimum difference increases. Ideally, the estimated measurement variability should be determined on the basis of preliminary data collected from a similar study population. A review of the literature can also provide estimates of this parameter. If preliminary data are not available, this parameter may have to be estimated on the basis of subjective experience, or a range of values may be assumed. A separate estimate of measurement variability is not required when the measurement being compared is a proportion (in contrast to a mean), because the SD is mathematically derived from the proportion.

### Statistical Power

This parameter is the power that is desired from the study. As power is increased, sample size increases. While high power is always desirable, there is an obvious trade-off with the number of individuals that can feasibly be studied, given the usually fixed amount of time and resources available to conduct a study. In randomized controlled trials, the statistical power is customarily set to a number greater than or equal to 0.80, with many clinical trial experts now advocating a power of 0.90.

### Significance Criterion

This parameter is the maximum *P* value for which a difference is to be considered statistically significant. As the significance criterion is decreased (made more strict), the sample size needed to detect the minimum difference increases. The significance criterion is customarily set to .05.

### One- or Two-tailed Statistical Analysis

In a few cases, it may be known before the study that any difference between comparison groups is possible in only one direction. In such cases, use of a one-tailed statistical analysis, which would require a smaller sample size for detection of the minimum difference than would a two-tailed analysis, may be considered. The sample size of a one-tailed design with a given significance criterion—for example, $\alpha$—is equal to the sample size of a two-tailed design with a significance criterion of $2\alpha$, all other parameters being equal. Because of this simple relationship and because truly appropriate one-tailed analyses are rare, a two-tailed analysis is assumed in the remainder of this article.

## SAMPLE SIZES FOR COMPARATIVE RESEARCH STUDIES

With knowledge of the design parameters detailed in the previous section, the calculation of an appropriate sample size simply involves selecting an appropriate equation. For a study comparing two means, the equation for sample size (13) is

$$N = \frac{4\sigma^2(z_{crit} + z_{pwr})^2}{D^2}, \qquad (1)$$

where $N$ is the total sample size (the sum of the sizes of both comparison groups), $\sigma$ is the assumed SD of each group (assumed to be equal for both groups), the $z_{crit}$ value is that given in Table 1 for the desired significance criterion, the $z_{pwr}$ value is that given in Table 2 for the desired statistical power, and $D$ is the minimum expected difference between the two means. Both $z_{crit}$ and $z_{pwr}$ are cutoff points along the x axis of a standard normal probability distribution that demarcate probabilities matching the specified significance criterion and statistical power, respectively. The two groups that make up

**TABLE 1**
**Standard Normal Deviate ($z_{crit}$) Corresponding to Selected Significance Criteria and CIs**

| Significance Criterion* | $z_{crit}$ Value† |
|---|---|
| .01 (99) | 2.576 |
| .02 (98) | 2.326 |
| .05 (95) | 1.960 |
| .10 (90) | 1.645 |

* Numbers in parentheses are the probabilities (expressed as a percentage) associated with the corresponding CIs. Confidence probability is the probability associated with the corresponding CI.

† A stricter (smaller) significance criterion is associated with a larger $z_{crit}$ value. Values not shown in this table may be calculated in Excel version 97 (Microsoft, Redmond, Wash) by using the formula $z_{crit}$ = NORMSINV(1−(*P*/2)), where *P* is the significance criterion.

**TABLE 2**
**Standard Normal Deviate ($z_{pwr}$) Corresponding to Selected Statistical Powers**

| Statistical Power | $z_{pwr}$ Value* |
|---|---|
| .80 | 0.842 |
| .85 | 1.036 |
| .90 | 1.282 |
| .95 | 1.645 |

* A higher power is associated with a larger value for $z_{pwr}$. Values not shown in this table may be calculated in Excel version 97 (Microsoft, Redmond, Wash) by using the formula $z_{pwr}$ = NORMSINV(*power*). For calculating power, the inverse formula is *power* = NORMSDIST($z_{pwr}$), where $z_{pwr}$ is calculated from Equation (1) or Equation (2) by solving for $z_{pwr}$.

*N* are assumed to be equal in number, and it is assumed that two-tailed statistical analysis will be used. Note that *N* depends only on the difference between the two means; it does not depend on the magnitude of either one.

As an example, suppose a study is proposed to compare a renovascular procedure versus medical therapy in lowering the systolic blood pressure of patients with hypertension secondary to renal artery stenosis. On the basis of results of preliminary studies, the investigators estimate that the vascular procedure may help lower blood pressure by 20 mm Hg, while medical therapy may help lower blood pressure by only 10 mm Hg. On the basis of their clinical judgment, the investigators might also argue that the vascular procedure would have to be twice as effective as medical therapy to justify the higher cost and discomfort of the vascular procedure. On the basis of results of preliminary studies, the SD for blood pressure lowering is estimated to be 15 mm Hg. According to the normal distribution, this SD indicates an expectation that 95% of the patients in either group will experience a blood pressure lowering within 30 mm Hg (2 SDs) of the mean. A significance criterion of .05 and power of 0.80 are chosen. With these assumptions, $D = 20 - 10 = 10$ mm Hg, $\sigma = 15$ mm Hg, $z_{crit} = 1.960$ (from Table 1), and $z_{pwr} = 0.842$ (from Table 2). Equation (1) yields a sample size of $N = 70.6$. Therefore, a total of 70 patients (rounding *N* to the nearest even number) should be enrolled in the study: 35 to undergo the vascular procedure and 35 to receive medical therapy.

For a study in which two proportions are compared with a $\chi^2$ test or a *z* test, which is based on the normal approximation to the binomial distribution, the equation for sample size (14) is

$$N = 2 \cdot [z_{crit}\sqrt{2\bar{p}(1-\bar{p})} + z_{pwr}\sqrt{p_1(1-p_1)+p_2(1-p_2)}]^2/D^2, \quad (2)$$

where $p_1$ and $p_2$ are pre-study estimates of the two proportions to be compared, $D = |p_1 - p_2|$ (ie, the minimum expected difference), $\bar{p} = (p_1 + p_2)/2$, and *N*, $z_{crit}$, and $z_{pwr}$ are defined as they are for Equation (1). The two groups comprising *N* are assumed to be equal in number, and it is assumed that two-tailed statistical analysis will be used. Note that in this case, *N* depends not only on the difference between the two proportions but also on the magnitude of the proportions themselves. Therefore, Equation (2) requires the investigator to estimate $p_1$ and $p_2$, as well as their difference, before performing the study. However, Equation (2) does not require an independent estimate of SD because it is calculated from $p_1$ and $p_2$ within the equation.

As an example, suppose a standard diagnostic procedure has an accuracy of 80% for the diagnosis of a certain disease. A study is proposed to evaluate a new diagnostic procedure that may have greater accuracy. On the basis of their experience, the investigators decide that the new procedure would have to be at least 90% accurate to be considered significantly better than the standard procedure. A significance criterion of .05 and a power of 0.90 are chosen. With these assumptions, $p_1 = 0.80$, $p_2 = 0.90$, $D = 0.10$, $\bar{p} = 0.85$, $z_{crit} = 1.960$, and $z_{pwr} = 0.842$. Equation (2) yields a sample size of $N = 398$. Therefore, a total of 398 patients should be enrolled: 199 to undergo the standard diagnostic procedure and 199 to undergo the new one.

## SAMPLE SIZES FOR DESCRIPTIVE STUDIES

Not all research studies involve the comparison of two groups. The purpose of many studies is simply to describe, with means or proportions, one or more characteristics in one particular group. In these types of studies, known as descriptive studies, sample size is important because it affects how precise the observed means or proportions are expected to be. In the case of a descriptive study, the minimum expected difference reflects the difference between the upper and lower limit of an expected *confidence interval*, which is described with a percentage. For example, a 95% CI indicates the range in which 95% of results would fall if a study were to be repeated an infinite number of times, with each repetition including the number of individuals specified by the sample size.

In studies designed to estimate a mean, the equation for sample size (2,15) is

$$N = \frac{4\sigma^2(z_{crit})^2}{D^2}, \quad (3)$$

where *N* is the sample size of the single study group, $\sigma$ is the assumed SD for the group, the $z_{crit}$ value is that given in Table 1, and *D* is the total width of the expected CI. Note that Equation (3) does not depend on statistical power because this concept only applies to statistical comparisons.

As an example, suppose a fetal sonographer wants to determine the mean fetal crown-rump length in a group of pregnancies. The sonographer would like the limits of the 95% confidence interval to be no more than 1 mm above or 1 mm below the mean crown-rump length of the group. From previous studies, it is known that the SD for the measurement is 3 mm. Based on these assumptions, $D = 2$ mm, $\sigma = 3$ mm, and $z_{crit} = 1.960$ (from Table 1). Equation (3) yields a sample size of $N = 35$. Therefore, 35 fetuses should be examined in the study.

In studies designed to measure a characteristic in terms of a proportion, the equation for sample size (2,15) is

$$N = \frac{4(z_{crit})^2 p(1-p)}{D^2}, \quad (4)$$

where *p* is a pre-study estimate of the proportion to be measured, and *N*, $z_{crit}$, and *D* are defined as they are for Equa-

tion (3). Like Equation (2), Equation (4) depends not only on the width of the expected CI but also on the magnitude of the proportion itself. Also like Equation (2), Equation (4) does not require an independent estimate of SD because it is calculated from $p$ within the equation.

As an example, suppose an investigator would like to determine the accuracy of a diagnostic test with a 95% CI of ±10%. Suppose that, on the basis of results of preliminary studies, the estimated accuracy is 80%. With these assumptions, $D = 0.20$, $p = 0.80$, and $z_{crit} = 1.960$. Equation (4) yields a sample size of $N = 61$. Therefore, 61 patients should be examined in the study.

## MINIMIZING THE SAMPLE SIZE

Now that we understand how to calculate sample size, what if the sample size we calculate is too large to be feasibly studied? Browner et al (16) list a number of strategies for minimizing the sample size. These strategies are briefly discussed in the following paragraphs.

### Use Continuous Measurements Instead of Categories

Because a radiologic diagnosis is often expressed in terms of a binary result, such as the presence or absence of a disease, it is natural to convert continuous measurements into categories. For example, the size of a lesion might be encoded as "small" or "large." For a sample of fixed size, the use of the actual measurement rather than the proportion in each category yields more power. This is because statistical tests that incorporate the use of continuous values are mathematically more powerful than those used for proportions, given the same sample size.

### Use More Precise Measurements

For studies in which Equation (1) or Equation (2) applies, any way to increase the precision (decrease the variability) of the measurement process should be sought. For some types of research, precision can be increased by simply repeating the measurement. More complex equations are necessary for studies involving repeated measurements in the same individuals (17), but the basic principles are similar.

### Use Paired Measurements

Statistical tests like the paired $t$ test are mathematically more powerful for a given sample size than are unpaired tests because in paired tests, each measurement is matched with its own control. For example, instead of comparing the average lesion size in a group of treated patients with that in a control group, measuring the change in lesion size in each patient after treatment allows each patient to serve as his or her own control and yields more statistical power. Equation (1) can still be used in this case. $D$ represents the expected change in the measurement, and $\sigma$ is the expected SD of this change. The additional power and reduction in sample size are due to the SD being smaller for changes within individuals than for overall differences between groups of individuals.

### Use Unequal Group Sizes

Equations (1) and (2) involve the assumption that the comparison groups are equal in size. Although it is statistically most efficient if the two groups are equal in size, benefit is still gained by studying more individuals, even if the additional individuals all belong to one of the groups. For example, it may be feasible to recruit additional individuals into the control group even if it is difficult to recruit more individuals into the noncontrol group. More complex equations are necessary for calculating sample sizes when comparing means (13) and proportions (18) of unequal group sizes.

### Expand the Minimum Expected Difference

Perhaps the minimum expected difference that has been specified is unnecessarily small, and a larger expected difference could be justified, especially if the planned study is a preliminary one. The results of a preliminary study could be used to justify a more ambitious follow-up study of a larger number of individuals and a smaller minimum difference.

## DISCUSSION

The formulation of Equations (1–4) involves two statistical assumptions which should be kept in mind when these equations are applied to a particular study. First, it is assumed that the selection of individuals is random and unbiased. The decision to include an individual in the study cannot depend on whether or not that individual has the characteristic or outcome being studied. Second, in studies in which a mean is calculated from measurements of individuals, the measurements are assumed to be normally distributed. Both of these assumptions are required not only by the sample size calculation method, but also by the statistical tests themselves (such as the $t$ test). The situations in which Equations (1–4) are appropriate all involve parametric statistics. Different methods for determining sample size are required for nonparametric statistics such as the Wilcoxon rank sum test.

Equations for calculating sample size, such as Equations (1) and (2), also provide a method for determining statistical power corresponding to a given sample size. To calculate power, solve for $z_{pwr}$ in the equation corresponding to the design of the study. The power can be then determined by referring to Table 2. In this way, an "observed power" can be calculated after a study has been completed, where the observed difference is used in place of the minimum expected difference. This calculation is known as retrospective power analysis and is sometimes used to aid in the interpretation of the statistical results of a study. However, retrospective power analysis is controversial because it can be shown that observed power is completely determined by the $P$ value and therefore cannot add any additional information to its interpretation (19). Power calculations are most appropriate when they incorporate a minimum difference that is stated prospectively.

The accuracy of sample size calculations obviously depends on the accuracy of the estimates of the parameters used in the calculations. Therefore, these calculations should always be considered estimates of an absolute minimum. It is usually prudent for the investigator to plan to include more than the minimum number of individuals in a study to compensate for loss during follow-up or other causes of attrition.

Sample size is best considered early in the planning of a study, when modifications in study design can still be made. Attention to sample size will hopefully result in a more meaningful study whose results will eventually receive a high priority for publication.

**References**

1. Pagano M, Gauvreau K. Principles of biostatistics. 2nd ed. Pacific Grove, Calif: Duxbury, 2000; 246–249, 330–331.
2. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7th ed. New York, NY: Wiley, 1999; 180–185, 268–270.
3. Altman DG. Practical statistics for medical research. London, England: Chapman & Hall, 1991.
4. Bond J. Power calculator. Available at: *http://calculators.stat.ucla.edu/powercalc/*. Accessed March 11, 2003.

5. Uitenbroek DG. Sample size: SISA—simple interactive statistical analysis. Available at: *http://home.clara.net/sisa/samsize.htm*. Accessed March 3, 2003.
6. Lenth R. Java applets for power and sample size. Available at: *www.stat.uiowa.edu/~rlenth/Power/index.html*. Accessed March 3, 2003.
7. NCSS Statistical Software. PASS 2002. Available at: *www.ncss.com/pass.html*. Accessed March 3, 2003.
8. SPSS. SamplePower. Available at: *www.spss.com/SPSSBI/SamplePower/*. Accessed March 3, 2003.
9. Statistical Solutions. nQuery Advisor. Available at: *www.statsolusa.com/nquery/nquery.htm*. Accessed March 3, 2003.
10. Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987; 257:2459–2463.
11. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994; 272:122–124.
12. 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.
13. Rosner B. Fundamentals of biostatistics. 5th ed. Pacific Grove, Calif: Duxbury, 2000; 308.
14. Feinstein AR. Principles of medical statistics. Boca Raton, Fla: CRC, 2002; 503.
15. Snedecor GW, Cochran WG. Statistical methods. 8th ed. Ames, Iowa: Iowa State University Press, 1989; 52, 439.
16. Browner WS, Newman TB, Cummings SR, Hulley SB. Estimating sample size and power. In: Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. Designing clinical research: an epidemiologic approach. 2nd ed. Philadelphia, Pa: Lippincott Williams & Wilkins, 2001; 65–84.
17. Frison L, Pocock S. Repeated measurements in clinical trials: analysis using mean summary statistics and its implications for design. Stat Med 1992; 11:1685–1704.
18. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY: Wiley, 1981; 45.
19. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 2001; 55:19–24.