Sample Size Determination in Epidemiologic Studies

Janet D. Elashoff, Stanley Lemeshow

1.1 Introduction 561

1.2 One Group Designs, Inferences About Proportions 561

Confidence Intervals for a Single Population Proportion 562
Hypothesis Testing for a Single Population Proportion 564
Additional Considerations and References 567

1.3 Comparison of Two Independent Proportions 567

Study Designs, Parameters, Analysis Methods 567
Confidence Intervals for the Risk Difference 569
Confidence Interval for Relative Risk (Ratio) 571
Confidence Intervals for the Odds Ratio 572
Testing the Difference Between Two Proportions 574
Testing the Relative Risk 575
Testing the Odds Ratio 576
Additional Considerations and References 577

1.4 One Group Designs, Inferences About a Single Mean 578

Confidence Intervals for a Single Mean 578
Hypothesis Testing for a Single Population Mean 580

1.5 Comparison of Two Independent Means 582

Confidence Intervals for the Difference Between Two Means 582
Testing the Difference Between Two Means (Two-Sample t Test) 584
Additional Considerations and References 585

1.6 Logistic Regression Models 587

Single Dichotomous Covariate 587

Single Continuous Covariate 588
Adjusting Sample Size for Inclusion of k Prior Covariates (Variance Inflation Factor) 589
Assessing the Adequacy of Data Already Collected 590

1.7 Practical Issues in Sample Size Choice 591
1.8 Conclusions 592
References 593

Introduction

1.1

When planning a research project an epidemiologist must consider how many subjects should be studied. While factors such as available budget certainly present constraints on the maximum number of subjects that might actually be included in a study, statistical considerations are extremely important. To address the statistical questions about appropriate sample size, the researcher must first specify the study design, the nature of the outcome variable, the aims of the study, the planned analysis method, and the expected results of the study. Is the goal of the study to distinguish between hypotheses about the value of a parameter or function of parameters, or is the goal to provide a confidence interval estimate of a parameter such as the odds ratio or relative risk?

This chapter is organized as follows. We introduce the issue of how to choose sample size for estimation of a parameter or for a hypothesis test regarding a parameter in the context of one-sample studies in which it is desired to estimate or test a population proportion. We continue on to two-sample studies involving comparisons between two proportions, and one and two-sample studies involving estimation or testing of population means. We conclude with a section on sample size for logistic regression.

In this chapter we will provide a brief introduction to power and sample size computation and only address sample size issues for a few of the procedures that are most commonly used in epidemiologic research. However, we do hope that the reader will gain a sense for what one can accomplish by planning a study with appropriate attention to sample size considerations.

A focus on sample size considerations when the study is first being planned is critical for the ultimate likelihood that a study proposal is accepted for funding and that the final manuscript will be accepted for publication. To ignore the issue of sample size would greatly increase the likelihood of embarking on a costly and time-consuming epidemiologic study with little likelihood of finding any definitive results.

One Group Designs, Inferences About Proportions

1.2

The simplest study design is one in which interest focuses on results for a single group. One is often interested in making inferences about the value of a population proportion. In this section we will illustrate how to choose sample size for the following examples:

Example 1 . A district medical officer seeks to estimate the proportion of children in the district receiving appropriate childhood vaccinations. Assuming a simple random sample is to be selected from a community, how many children must be studied if the resulting estimate is to fall within 10 percentage points of the true proportion with 95% confidence? ♦

Example 2 . Consider the information given in Example 1, only this time we will determine the sample size necessary to estimate the proportion vaccinated in the population to within 10% (not 10 percentage points) of the true value. ♦

Example 3 . During a virulent outbreak of neonatal tetanus, health workers wish to determine whether the rate is decreasing after a period during which it had risen to a level of 150 cases per thousand live births. What sample size is necessary to test the null hypothesis that the population proportion is 0.15 at the 0.05 level if it is desired to have a 90% probability of detecting a decrease to a rate of 100 per thousand if that were the true proportion? ♦

The first two examples involve estimation and confidence intervals while the third involves a statistical hypothesis test.

The usual model underlying testing or estimation of a population proportion assumes that the design involves a simple independent random sample from a population in which the probability of a "success" is constant. The distribution of the number of successes in a sample of size n with a true underlying proportion of successes denoted by π is given by the binomial distribution. However, formulas are simplified when power and sample size determinations are made on the basis of using the normal approximation to the binomial.

The sampling distribution of the sample proportion "p" is approximately normal with mean of π (the expected value of p, $E(p) = \pi$) and variance of p, $\text{Var}(p) = \pi(1-\pi)/n$; the standard deviation is $\sqrt{\pi(1-\pi)/n}$.

We begin by discussing sample size determination for estimation (the confidence interval approach) and then turn to sample size determination for hypothesis testing problems.

1.2.1 Confidence Intervals for a Single Population Proportion

Two-sided $100(1-\alpha)\%$ confidence intervals for a parameter, θ, based on using the normal approximation can be stated in general as:

$$\widehat{\theta} \pm z_{1-\alpha/2}\widehat{SE}\left(\widehat{\theta}\right) \ , \tag{1.1}$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the normal (or Gaussian) distribution. For the commonly used two-sided 95% confidence interval, $z_{1-\alpha/2} = 1.96$. The $100(1-\alpha)\%$ confidence interval for π based on the estimated proportion, p, is given by

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \,. \tag{1.2}$$

Letting, ω be the half-width of the confidence interval for the expected true value π, we have

$$\omega = z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \,. \tag{1.3}$$

The sample size necessary to achieve a confidence interval of width ω is given by

$$n = \left(\frac{z_{1-\alpha/2}}{\omega}\right)^2 \left[\pi(1-\pi)\right] \,. \tag{1.4}$$

Returning to Example 1, we begin by assuming that the rate of vaccinated children is expected to be about 75%. We would then set $\pi = 0.75$, $\omega = 0.10$ and $z_{1-\alpha/2} = 1.96$. From (1.4) we find that $n = 72.03$. Note that for sample size calculations we round up. We conclude that to estimate the expected population proportion to within ± 0.10, a sample of 73 children would be required.

If we don't really know what rate to expect we can make use of the fact that n will be largest for $\pi = 0.50$ and use this value to solve for n. For Example 1 we require a sample size of 97 to be sure that the confidence interval width will be no wider than plus or minus 10 percentage points no matter what the observed proportion is.

Table 1.1 presents the required sample sizes for selected values of π and ω.

Table 1.1. Sample size for 95% two-sided confidence interval for a proportion (using the normal approximation) to have expected width, ω

	ω	
π	± 0.05	± 0.10
0.50	385	97
0.25	289	73
0.10	139	35

Proceeding to Example 2, we consider the information given in Example 1, only this time we will determine the sample size necessary to estimate the proportion vaccinated in the population to within 10% (not 10 percentage points) of the true value.

Let θ be the unknown population parameter as before and let $\widehat{\theta}$ be the estimate of θ. Let ε, the desired precision, be defined as:

$$\varepsilon = \frac{\left|\widehat{\theta} - \theta\right|}{\theta} \,.$$

In the present example, based on the confidence limits using the normal approximation to the distribution of p, it follows that

$$\left|p - \pi\right| = z_{1-\alpha/2}\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}$$

and, dividing both sides by π, an expression similar to the one presented above for ε is obtained. That is,

$$\varepsilon = \frac{\left|p - \pi\right|}{\pi} = z_{1-\alpha/2}\frac{\sqrt{1-\pi}}{\sqrt{n\pi}}$$

and squaring both sides and solving for n gives:

$$n = z^2_{1-\alpha/2}\frac{1-\pi}{\varepsilon^2\pi} \ . \tag{1.5}$$

Assuming $\pi = 0.75$, we would find that a sample size of 129 would be required to assure that the 95% confidence interval would be within 10% of the true value.

1.2.2 Hypothesis Testing for a Single Population Proportion

Suppose we would like to test a null hypothesis about the value of the population proportion

$$H_0 : \pi = \pi_0$$

versus the one-sided alternative hypothesis

$$H_\text{a} : \pi > \pi_0 \ .$$

Statistical hypothesis testing involves balancing the two types of errors that can be made. Type I error is defined as the error of rejecting the null hypothesis when it is in fact true. We denote the probability of making a Type I error as "α"; a commonly used choice for α is 0.05. The critical value of the test statistic is then chosen so that the probability of rejecting the null hypothesis when it is true will be α.

To choose the necessary sample size, we need to address Type II error as well. A Type II error is the error of failing to reject the null hypothesis when it is in fact false. To determine the probability of a Type II error (denoted by "β"), we must specify a particular value of interest for the alternative hypothesis, say, π_a. The probability of rejecting the null hypothesis when it is false is defined as the *power* of the test, $1 - \beta$. Typically, we require the power at the alternative of interest to be 80% or 90%.

Based on the normal approximation to the binomial, the test statistic for a test of the null hypothesis is given by

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \ .$$

To set the probability of a Type I error equal to α, we plan to reject the null hypothesis if $z > z_{1-\alpha}$. To choose n, we fix the probability that $z > z_{1-\alpha}$ if the population proportion equals π_a to be $1-\beta$. This may be represented graphically as shown in Fig. 1.1:

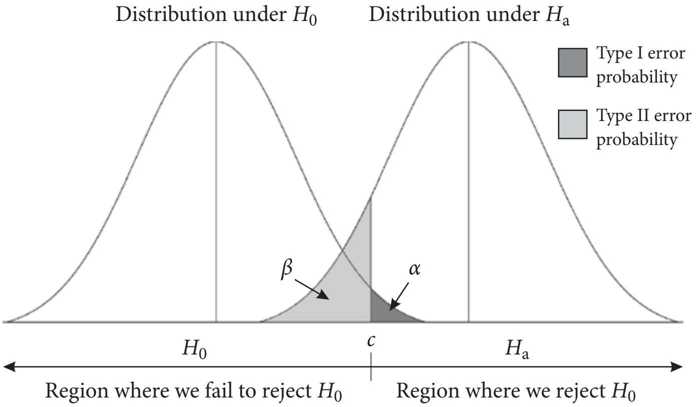

Figure 1.1. Sampling distributions for one-sample hypothesis test

In this figure the point "c" represents the upper $100\,\alpha$th percent point of the distribution of p for the sampling distribution centered at π_0 (i.e., the distribution which would result if the null hypothesis were true):

$$c = \pi_0 + z_{1-\alpha}\sqrt{\pi_0\left(1-\pi_0\right)/n}\ .$$

For the sampling distribution centered at π_a (i.e., the distribution which would result if the alternate hypothesis were true), "c" represents the lower $100\,\beta$th percent point of the distribution of p:

$$c = \pi_a + z_\beta\sqrt{\pi_a\left(1-\pi_a\right)/n}\ .$$

In order to find n we set the two expressions equal to each other. From this, it follows that:

$$\pi_0 + z_{1-\alpha}\sqrt{\pi_0\left(1-\pi_0\right)/n} = \pi_a + z_\beta\sqrt{\pi_a\left(1-\pi_a\right)/n}\ .$$

Noting that $z_{1-\beta} = -z_\beta$, we find

$$\pi_a - \pi_0 = \frac{\left\{ z_{1-\alpha}\sqrt{\pi_0\left(1-\pi_0\right)} + z_{1-\beta}\sqrt{\pi_a\left(1-\pi_a\right)} \right\}}{\sqrt{n}}$$

and, solving for n, we find that the necessary sample size, for this single sample hypothesis testing situation, is given by the formula:

$$n = \frac{\left\{z_{1-\alpha}\sqrt{\pi_0\left(1-\pi_0\right)} + z_{1-\beta}\sqrt{\pi_a\left(1-\pi_a\right)}\right\}^2}{\left(\pi_a - \pi_0\right)^2} \,. \tag{1.6}$$

Notice that as π_a gets further and further away from π_0, the necessary sample size decreases.

To illustrate, we return to Example 3 in which we wish to test the null hypothesis that $\pi = 0.15$ at the one-sided 5% level and have 90% power to detect a decrease to a rate of 0.10. Using (1.6), it follows that

$$n = \frac{\left\{1.645\sqrt{0.15(0.85)} + 1.282\sqrt{0.10(0.90)}\right\}^2}{(0.05)^2} = 377.90 \,.$$

Hence we see that a total sample size of 378 live births would be necessary.

To plan sample size for a two-sided test, we need only substitute $z_{1-\alpha/2}$ for $z_{1-\alpha}$ in (1.6) to obtain:

$$n = \frac{\left\{z_{1-\alpha/2}\sqrt{\pi_0\left(1-\pi_0\right)} + z_{1-\beta}\sqrt{\pi_a\left(1-\pi_a\right)}\right\}^2}{\left(\pi_a - \pi_0\right)^2} \,. \tag{1.7}$$

To have 90% power for a two-sided 5% level test for Example 3 would require a total of 471 subjects to detect the difference between the null hypothesis proportion, π_0, of 0.15 and the alternative proportion, π_a, of 0.10. Note that the sample size required to achieve 90% power for the specified alternative is larger when a two-sided 5% level test is planned than when a one-sided 5% level test is planned, so that the investigator needs to be clear as to whether the planned test is to be one-sided or two-sided when making sample size computations.

Table 1.2. Sample size for 0.05-level, two-sided test that the proportion equals π_0 versus the alternative π_a for specified levels of power (based on normal approximation)

		Power	
π_0	π_a	80%	90%
0.50	0.40	194	259
0.50	0.30	47	62
0.20	0.10	108	137
0.15	0.10	363	471
0.10	0.05	239	301

Table 1.2 presents the required sample sizes for selected values of π_0, π_a and power. For a two-sided test, unless the null hypothesis proportion equals 0.5, computed sample sizes for alternative proportions given by $\pi_{aL} = \pi_0 - \delta$ and

$\pi_{aU} = \pi_0 + \delta$ will differ; the larger estimate of sample size will be obtained for the alternative proportion closer to 0.5.

Additional Considerations and References 1.2.3

Good introductions to sample size computations for tests and confidence intervals for a single proportion can be found in Dixon and Massey (1983), Lemeshow et al. (1990), Fleiss (1981) and Lachin (1981). Books containing sample size tables are available (e.g. Machin and Campbell 1987; Machin et al. 1997; Lemeshow et al. 1990). Commercially available sample size software such as nQuery Advisor Release 6 (Elashoff 2005) can be used to compute sample size for confidence intervals or hypothesis tests (based on either the normal approximation or an exact binomial test) for a single proportion as well as for a wide variety of other sample size problems.

For values of π near 0 or 1 (or for small sample sizes), sample size methods involving a continuity correction (Fleiss et al. 1980), methods designed for rare events (e.g. Korn 1986; Louis 1981), or methods based on exact tests (Chernick and Liu 2002) may be preferable.

Note that an actual field survey is unlikely to be based on a simple random sample. As a result, the required sample size would go up by the amount of the "design effect" which is determined by the details of the actual sampling plan. The "design effect" is the ratio of the standard error of the estimated parameter under the study design to the standard error of the estimate under simple random sampling; a text on sample surveys should be consulted for details (see Levy and Lemeshow (1999)). For example, if a cluster sampling plan with a design effect of 2 were to be employed, the sample size computed using the above formulas would need to be doubled.

Comparison of Two Independent Proportions 1.3

Study Designs, Parameters, Analysis Methods 1.3.1

More sample size literature exists for the problem of comparing two independent proportions than for any other sample size problem. This has come about because there are several basic sampling schemes leading to problems of this type. There are different parameterizations of interest and a variety of test and estimation procedures that have been developed. Sample size formulations depend on the parameter of interest for testing or estimation as well as the specifics of the test or estimation procedure.

The basic study designs relevant to epidemiological studies are experimental trials, cohort studies, and case-control studies. We describe each study type briefly and give an example. The examples will be addressed in more detail in subsequent sections.

Experimental Trial. $2n$ subjects are recruited for a study; n are randomly assigned to group 1 and n to group 2. The intervention is applied according to the design. Subjects are followed for a fixed time and success-failure status is recorded. Experimental trials are usually randomized, often double blind, and always prospective. For example, patients with intestinal parasites are randomly assigned to receive either the standard drug or a new drug and followed to determine whether they respond favorably. The observed proportion responding favorably in group i is denoted by p_i and the true population proportion in group i by π_i.
Experimental trials are typically analyzed in terms of the difference in proportions, or the risk difference.

$$\text{Population risk difference} = \pi_1 - \pi_2 \tag{1.8}$$

$$\text{Estimated risk difference} = p_1 - p_2 \tag{1.9}$$

Cohort Study. n subjects are recruited from group 1 and n from group 2; subjects are followed for a fixed time and success-failure status is recorded. Cohort studies are typically prospective studies. For example, workers with asbestos exposure and workers in the same industry without asbestos exposure are followed for the development of lung disease.
Cohort studies may be analyzed in terms of the risk difference or in terms of the relative risk.

$$\text{Population relative risk} = RR = \pi_2/\pi_1 \tag{1.10}$$

$$\text{Estimated relative risk} = rr = p_2/p_1 \tag{1.11}$$

Referring to the example, π_1 denotes the true proportion of diseased workers in the unexposed group while π_2 denotes the true proportion of diseased workers in the exposed group, and p_1 and p_2 are the corresponding observed proportions.

Case-Control Studies. n subjects (cases) are recruited from among those who have developed a disease and n subjects (controls) are recruited from a similar group without the disease. Subjects from both groups are studied for the presence of a relevant exposure in their background. For example, tuberculosis (TB) cases and controls are assessed for whether they had been vaccinated with BCG (Bacillus Calmette-Guérin vaccine). Case-control studies are inherently retrospective studies and interest is focused on the odds ratio.

$$\text{Population odds ratio} = OR = \pi_2\left(1 - \pi_1\right)/\left(1 - \pi_2\right)\pi_1 \tag{1.12}$$

$$\text{Estimated odds ratio} = or = p_2\left(1 - p_1\right)/\left(1 - p_2\right)p_1 \tag{1.13}$$

Referring to the example, π_1 denotes the true proportion of vaccinated subjects among the controls while π_2 denotes the true proportion of vaccinated subjects among the TB cases, and p_1 and p_2 are the corresponding observed proportions.

We begin by discussing sample size determination for estimation (the confidence interval approach) and then turn to sample size determination for hypothesis testing problems.

Confidence Intervals for the Risk Difference 1.3.2

Example 4 . A pilot study with 20 subjects randomized to receive the standard drug to control intestinal parasites and 20 to receive a new drug found that 13 subjects (65%) receiving the standard drug responded favorably while 17 (85%) of the subjects receiving the new drug responded favorably.

Question 4a: Do these data establish that the new drug is better (lower limit of confidence interval is greater than zero) and, if not, might it still be enough better to warrant a larger clinical trial? We address this question with a confidence interval below.

Question 4b: What sample size would be required for the larger clinical trial? We address this question in the context of a confidence interval later in this section, and in the context of a hypothesis test in the following section. ♦

The estimated value of the risk difference, $\pi_1 - \pi_2$, is given by $p_1 - p_2$, the observed difference in proportions. The variance of $p_1 - p_2$ for independent proportions when the sample sizes, n, in each group are equal is:

$$\text{Var}\,(p_1 - p_2) = \frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{n}\,. \tag{1.14}$$

This formula is based on the assumption that the data come from independent random samples from the populations of interest. In population i, the probability of a success is a constant, π_i, and therefore the number of successes observed for each group has a binomial distribution with parameters n and π_i.

The standard error of this estimate, $p_1 - p_2$, is estimated by substituting the observed proportions for the true proportions and is given by

$$SE(p_1 - p_2) = \frac{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}{\sqrt{n}}\,. \tag{1.15}$$

Referring to the basic formula for a confidence interval based on the normal approximation given in (1.1), a two-sided 95% confidence interval for the difference in the proportions responding favorably to the new drug in comparison to the old drug is given by

$$0.85 - 0.65 \pm 1.96\frac{\sqrt{0.85(1-0.85) + 0.65(1-0.65)}}{\sqrt{20}}\,.$$

The limits are 0.20 ± 0.209 or −0.009 to 0.409, suggesting that although we cannot rule out a difference of zero the data indicate that the new drug might work markedly better than the standard.

The investigator wants to plan a definitive study to assess how much the success rates really do differ. What sample size would be necessary to obtain a confidence interval whose width is less than or equal to ±0.05?

We require that the confidence interval for $\pi_1 - \pi_2$, be $p_1 - p_2 \pm \omega$, where for Example 4, $\omega \leq 0.05$. To obtain a confidence interval width satisfying these conditions, we must have

$$z_{1-\alpha/2} \frac{\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}}{\sqrt{n}} \leq \omega \, .$$

Solving this equation for n, the sample size in each group, we obtain (1.16).

$$n = \frac{z_{1-\alpha/2}^2 \left[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)\right]}{\omega^2} \, . \tag{1.16}$$

For Example 4, an n per group of 546 would be required to obtain an expected 95% two-sided confidence interval width of approximately ±0.05 if we expect to see about the same proportions as we did in the pilot study.

Table 1.3 presents the sample size in each group necessary to obtain specified confidence interval widths for a few selected examples. This table should provide investigators with a quick idea of the order of magnitude of required sample sizes. Note that since the confidence interval width depends on the postulated proportions only through the terms $\pi_i\left(1-\pi_i\right)$, this table can also be used for proportions greater than 0.5.

If an investigator is a bit uncertain about what proportions to expect and wants to ensure that the confidence interval width is less than some specified amount $\pm\omega$ no matter what proportions are observed, we can use the fact that the confidence interval is widest when $\pi_1 = \pi_2 = 0.5$. In this case the sample size required for each group is

$$n \leq \frac{z_{1-\alpha/2}^2}{2\omega^2} \, . \tag{1.17}$$

Table 1.3. Sample size per group for 95% two-sided confidence interval (using normal approximation) for risk difference to have expected width, ω

		ω	
π_1	π_2	±0.05	±0.10
0.50	0.50	769	193
0.50	0.25	673	169
0.50	0.10	523	131
0.25	0.25	577	145
0.25	0.10	427	107
0.10	0.10	277	70

For a two-sided 95% confidence interval this becomes approximately $2/\omega^2$. For Example 4, the maximum sample size per group required for a confidence interval width of no more than ±0.05 is 769.

Confidence Interval for Relative Risk (Ratio)

1.3.3

Example 5 . Workers with asbestos exposure and workers in the same industry without asbestos exposure are followed for the development of lung disease. Suppose that disease occurs in 20% of the unexposed group, how large a sample would be needed in each of the exposed and unexposed study groups to estimate the relative risk to within 10% of the true value with 95% confidence assuming that the relative risk is approximately 1.75? ♦

For this purpose we define group 1 as the unexposed group and group 2 as the exposed group. The estimate of the relative risk (cf. Chap. I.2 of this handbook) is

$$\widehat{RR} = rr = p_2/p_1 \ .$$

Since we are dealing with a ratio, which can be expected to have a skewed distribution with a log-normal shape, we need to take logs to normalize the distribution so that the normal approximation can be used to construct the confidence interval.

We obtain the standard deviation for the estimate for the case where the sample sizes in the two groups are equal by using the approximation

$$\text{Var}\left(\ln(rr)\right) \approx \frac{1-\pi_1}{n\pi_1} + \frac{1-\pi_2}{n\pi_2} \ . \qquad (1.18)$$

The estimated standard deviation is obtained by substituting the estimated proportions for the population proportions and taking the square root.

The $100(1-\alpha)\%$ confidence limits for $\ln(RR)$ are given by $\ln(rr) \pm \omega$ where

$$\omega = z_{1-\alpha/2}\widehat{SE}\left(\ln(rr)\right) = z_{1-\alpha/2}\sqrt{\frac{1-\pi_1}{n\pi_1} + \frac{1-\pi_2}{n\pi_2}} \ .$$

Then the confidence limits for RR are given by $\exp\left(\ln\left(rr_L\right)\right)$ and $\exp\left(\ln\left(rr_U\right)\right)$ where $\ln\left(rr_L\right)$ and $\ln\left(rr_U\right)$ are the lower and upper confidence limits for $\ln(RR)$.

To choose the sample size necessary to obtain a confidence interval of a desired width for $\ln(RR)$, we could simply specify ω and solve for n.

$$n = \frac{z_{1-\alpha/2}^2\left[(1-\pi_1)/\pi_1 + (1-\pi_2)/\pi_2\right]}{\omega^2} \ . \qquad (1.19)$$

Alternatively, an investigator may wish to specify the width in terms of how close the limits are to RR. For example, suppose that we are thinking in terms of values of $RR > 1$, and that we want the difference between RR and RR_L to be no greater than

εRR; that is, we set $RR - RR_L = \varepsilon RR$ which we rearrange to get $RR(1-\varepsilon) = RR_L$. Then, taking logs, we have

$$\ln(RR) + \ln(1-\varepsilon) = \ln\left(RR_L\right)$$

and

$$\ln(RR) - \ln\left(RR_L\right) = -\ln(1-\varepsilon) = \omega$$

so

$$\omega = z_{1-\alpha/2}\sqrt{\frac{1-\pi_1}{n\pi_1} + \frac{1-\pi_2}{n\pi_2}} = -\ln(1-\varepsilon)\ .$$

Then to find the necessary sample size for each group, we solve for n to obtain

$$n = \frac{z^2_{1-\alpha/2}\left[(1-\pi_1)/\pi_1 + (1-\pi_2)/\pi_2\right]}{[\ln(1-\varepsilon)]^2}\ . \tag{1.20}$$

A version of this, which substitutes the expected RR for π_2, is

$$n = \frac{z^2_{1-\alpha/2}\left[(1+RR)/(RR\pi_1) - 2\right]}{[\ln(1-\varepsilon)]^2}\ . \tag{1.21}$$

Returning to Example 5, the expected $RR = 1.75$, $\pi_1 = 0.20$, and we have requested that the lower limit of the confidence interval for RR be within 10% of the true value of RR. Therefore $\varepsilon = 0.1$, $1-\varepsilon = 0.9$ and the required sample size would be 2027 per group or 4054 total.

Table 1.4 presents the sample size in each group necessary to obtain specified confidence interval widths for a few selected examples.

Table 1.4. Sample size per group for 95% two-sided confidence interval for the relative risk to have lower limit $(1-\varepsilon)RR$

		ε	
RR	π_1	0.10	0.20
1.25	0.20	2423	540
1.50	0.20	2192	489
1.75	0.20	2027	452
2.00	0.20	1904	424
1.25	0.40	866	193

1.3.4 Confidence Intervals for the Odds Ratio

Example 6. The efficacy of BCG vaccine in preventing childhood tuberculosis is in doubt and a study is designed to compare the immunization coverage rates in a group of tuberculosis cases compared to a group of controls.

Available information indicated that roughly 30% of controls are not vaccinated and we wish to estimate the odds ratio to within 20% of the true value. It is believed that the odds ratio is likely to be about 2.0. ♦

For problems involving estimation of the odds ratio (cf. Chap. I.2 of this handbook) we let group 1 denote the controls and group 2 denote the cases. Our estimate of the odds ratio is

$$or = \frac{p_2\left(1 - p_1\right)}{\left(1 - p_2\right) p_1} \,.$$

Since we are dealing with a ratio we need to take logs so that the normal approximation can be used to construct the confidence interval.

We obtain the standard deviation for the estimate for the case where the sample sizes in the two groups are equal by using the approximation

$$\text{Var}\,\left(\ln(or)\right) \approx \frac{1}{n\pi_1(1-\pi_1)} + \frac{1}{n\pi_2(1-\pi_2)} \,. \tag{1.22}$$

The estimated standard deviation is obtained by substituting the estimated proportions for the population proportions and taking the square root.

To obtain a $100(1-\alpha)\%$ confidence interval for $\ln(OR)$ of width ω where $\omega = z_{1-\alpha/2} SE\left(\ln(or)\right)$ when the sample sizes in the two groups are equal we require a sample size per group of

$$n = \frac{z^2_{1-\alpha/2}\left[1/\left(\pi_2(1-\pi_2)\right) + 1/\left(\pi_1(1-\pi_1)\right)\right]}{\omega^2} \,. \tag{1.23}$$

In situations where we assume that the odds ratio is greater than 1.0, to specify that the lower limit of the confidence interval be no less than $(1-\varepsilon)OR$, we would set $\omega = -\ln(1-\varepsilon)$ as we did in the previous section for the relative risk. We then obtain

$$n = \frac{z^2_{1-\alpha/2}\left[1/\left(\pi_2(1-\pi_2)\right) + 1/\left(\pi_1(1-\pi_1)\right)\right]}{[\ln(1-\varepsilon)]^2} \,. \tag{1.24}$$

Solving for π_2 using (1.12), we have

$$\pi_2 = \frac{OR\pi_1}{OR\pi_1 + (1-\pi_1)}$$

and we can obtain sample size expressed in terms of π_1 and OR.

$$n = z^2_{1-\alpha/2}\left[\frac{OR + (1 - \pi_1 + OR\pi_1)^2}{\pi_1(1-\pi_1)OR[\ln(1-\varepsilon)]^2}\right] \,. \tag{1.25}$$

For Example 6, we have $OR = 2$, $\pi_1 = 0.30$, $(\pi_2 = 0.462)$ and $(1-\varepsilon) = 0.8$, so we need 678 subjects per group.

Table 1.5 presents the sample size in each group necessary to obtain specified confidence interval widths for OR for a few selected examples.

Table 1.5. Sample size per group for 95% two-sided confidence interval for *OR* to have lower limit $(1-\varepsilon)OR$

		ε	
OR	π_1	0.10	0.20
1.25	0.30	3171	708
1.50	0.30	3101	692
1.75	0.30	3061	683
2.00	0.30	3040	678
1.25	0.50	2786	621

1.3.5 Testing the Difference Between Two Proportions

The goal is to test

$$H_0 : \pi_1 = \pi_2 \quad \text{versus} \quad H_1 : \pi_1 \neq \pi_2 \,.$$

If it can be assumed that the samples of size n from both groups arise from independent binomial distributions, the test for H_0 can be performed using the normal approximation to the binomial.

The test statistic is

$$z = \frac{\sqrt{n}\left(p_1 - p_2\right)}{\sqrt{2\bar{p}\left(1-\bar{p}\right)}} \,, \tag{1.26}$$

where $z \sim N(0,1)$, i.e. z is normally distributed with mean 0 and variance 1, and where, in the general case with unequal sample sizes in the two groups,

$$\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \,,$$

whereas for equal sample sizes

$$\bar{p} = \frac{p_1 + p_2}{2} \,.$$

(Note that the two-sided z test given by (1.26) is algebraically equivalent to the standard χ^2 test.)

The sample size in each group required for a two-sided $100(1-\alpha)\%$ test to have power $1-\beta$ is

$$n = \frac{\left[z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})} + z_{1-\beta}\sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}\right]^2}{\left(\pi_1 - \pi_2\right)^2} \tag{1.27}$$

and $\bar{\pi}$ is defined by analogy with $\bar{p}$.

Example 7 . Typically, the outcome measure for placebo controlled double-blind trials for acute duodenal ulcer healing is the proportion of patients whose ulcer has healed by four weeks as ascertained by endoscopy. The healing rate for the placebo group is typically about 40%. H_2-blocking active drugs usually result in 70% healed. The investigator wishes to evaluate a new drug with the expectation of seeking FDA (US Food and Drug Administration) approval; the results will be assessed by comparing observed proportions healed using the χ^2 test at the two-sided 5% significance level. Such trials are expensive to mount so that if the new drug is as effective as those currently approved, the investigator wants a 90% chance that the trial will yield a significant result. ♦

Using (1.27) for a two-sided 5% test, a sample size of 56 patients per group or a total sample size of 112 patients would be required to achieve 90% power.

Table 1.6 presents the sample size in each group necessary for a 5% two-sided χ^2 test comparing two independent proportions to have specified power for a few selected examples.

Table 1.6. Sample size per group for 5% two-sided χ^2 test for the difference between two independent proportions to have specified power

		Power	
π_1	π_2	80%	90%
0.10	0.05	435	582
0.25	0.10	100	133
0.50	0.25	58	77
0.50	0.10	20	26

Testing the Relative Risk 1.3.6

In a cohort study, where we want to focus attention on a test of the relative risk

$$H_0 : RR = \frac{\pi_2}{\pi_1} = 1 ,$$

the large sample test for this null hypothesis is the same as for the null hypothesis that the difference in proportions is zero and therefore the sample size formulas are the same. If we substitute RR into (1.27) we obtain

$$n = \frac{\left[z_{1-\alpha/2}\sqrt{(1+RR)[1-\pi_1(1+RR)/2]} + z_{1-\beta}\sqrt{[1+RR-\pi_1(1+RR^2)]}\right]^2}{\pi_1(1-RR)^2} . \tag{1.28}$$

Example 8. Two competing therapies for a particular cancer are to be evaluated by the cohort study strategy in a multi-center clinical trial. Patients are randomized to either treatment A or B and are followed for recurrence of disease for five years following treatment. How many patients should be studied in each of the two arms of the trial in order to have 90% power to reject $H_0 : RR = 1$ in favor of the alternative $RR = 0.5$, if the test is to be performed at the two-sided $\alpha = 0.05$ level and it is assumed that $\pi_1 = 0.35$? ♦

For Example 8, we substitute $\pi_1 = 0.35$ and $RR = 0.5$ into (1.28) and find that the required sample size per group would be 131 or 262 total. Or we could have noted that $\pi_2 = 0.175$ and used (1.27).

Table 1.7 presents the sample size in each group necessary for a 5% two-sided normal approximation test of the null hypothesis that the relative risk is 1.0 to have specified power for a few selected examples.

Table 1.7. Sample size per group for 5% two-sided test that the relative risk equals 1 to have specified power

		Power	
RR	π_1	80%	90%
1.25	0.20	1094	1464
1.50	0.20	294	392
1.75	0.20	138	185
2.00	0.20	82	109
1.25	0.40	388	519

1.3.7 Testing the Odds Ratio

The null hypothesis that the odds ratio equals 1.0 can be tested using (1.26) as for the test of difference in proportions. Sample size formulas can be modified to be based on π_2 and *OR* by algebraic substitution in (1.27) if desired, however formulas are simpler if we use (1.12) to solve for the other proportion and use (1.27) directly.

Example 9. The efficacy of BCG vaccine in preventing childhood tuberculosis is in doubt and a study is designed to compare the immunization coverage rates in a group of tuberculosis cases compared to a group of controls. Available information indicates that roughly 30% of the controls are not vaccinated, and we wish to have an 80% chance of detecting whether the odds ratio is significantly different from 1 at the 5% level. If an odds ratio of 2 would be considered an important difference between the two groups, how large a sample should be included in each study group? ♦

For Example 9, $\pi_1 = 0.3$ and $OR = 2$ and thus $\pi_2 = 0.462$; so using (1.27) we find that to obtain 80% power for a two-sided 5% level test would require 141 subjects per group or 282 total.

Table 1.8 presents the sample size in each group necessary to obtain specified power for tests of $OR = 1$ for a few selected examples.

Table 1.8. Sample size per group for 5% two-sided test of $OR = 1$ to have specified power

		Power	
OR	π_1	80%	90%
1.25	0.30	1442	1930
1.50	0.30	425	569
1.75	0.30	219	293
2.00	0.30	141	188
1.25	0.50	1267	1695

Additional Considerations and References 1.3.8

Good introductions to sample size computations for tests and confidence intervals for comparing two independent proportions can be found in Dixon and Massey (1983), Lemeshow et al. (1990), Fleiss (1981) and Lachin (1981). Books containing sample size tables are available (e.g. Machin and Campbell 1987; Machin et al. 1997; Lemeshow et al. 1990). Commercially available sample size software such as nQuery Advisor Release 6 (Elashoff 2005) can be used to compute sample size (or width) for confidence intervals and sample size (or power) for hypothesis tests for the two proportion case (based on either the normal approximation, continuity corrected normal approximation or Fisher's exact test) as well as for a wide variety of other sample size problems.

For values of π near 0 or 1 (or for small sample sizes), sample size methods involving a continuity correction (Fleiss et al. 1980), or methods based on exact tests (Chernick and Liu 2002) may be preferable.

When plans call for the sample sizes in the two groups to be unequal, the formulas for sample size and power must incorporate the expected ratio of the sample sizes, see references above. Generally for the same total sample size, power will tend to be higher and confidence interval widths narrower when sample sizes are equal; for comparisons of proportions, total sample size will depend on whether the proportion closer to 0.5 has the larger or the smaller sample size.

Note that the sample size methods discussed above do not apply to correlation/agreement/repeated measures (or pair-matched case-control) studies in which N subjects are recruited and each subject is measured by two different raters, or is studied under two different treatments in a cross-over design. These designs cannot be analyzed using the methods described for independent proportions; for example, sample size computations for the difference between two correlated proportions are based on the McNemar test (Lachin 1992).

1.4 One Group Designs, Inferences About a Single Mean

We turn to consideration of continuous outcomes and to inferences about the population mean. We denote the true but unknown mean in the population by μ and assume that the standard deviation for the population is given by σ. For a random sample of size n from a population with a normal (Gaussian) distribution, the distribution of the observed sample mean, $\bar{x}$, will also be normal with mean μ and standard deviation (also referred to as the standard error) given by $SE(\bar{x}) = \sigma/\sqrt{n}$. By the central limit theorem, the sampling distribution of the sample mean can usually be expected to be approximately normal for sample sizes of 30 or above even when the underlying population distribution is not normal.

1.4.1 Confidence Intervals for a Single Mean

Example 10 . Suppose an estimate is desired of the average retail price of twenty tablets of a commonly used tranquilizer. A random sample of retail pharmacies is to be selected. The estimate is required to be within 10 cents of the true average price with 95% confidence. Based on a small pilot study, the standard deviation in price, σ, can be estimated as 85 cents. How many pharmacies should be randomly selected? ♦

Using the normal approximation, the two-sided $100(1-\alpha)\%$ confidence interval for the true mean, μ, for the case where the standard deviation is known, is given by

$$\bar{x} \pm z_{1-\alpha/2}\sigma/\sqrt{n} \,. \tag{1.29}$$

So the sample size required to obtain a confidence interval of width ω is

$$n = \frac{z^2_{1-\alpha/2}\sigma^2}{\omega^2} \,. \tag{1.30}$$

For Example 10, expressing costs in dollars,

$$n = \frac{(1.96)^2(0.85)^2}{(0.10)^2} = 277.6 \,.$$

Therefore a sample size of 278 pharmacies should be selected.

We should note however that usually the standard deviation must be estimated from the sample. Then, the actual confidence interval for a sample mean would be given by

$$\bar{x} \pm t_{n-1,1-\alpha/2}s/\sqrt{n} \,, \tag{1.31}$$

where s is the observed standard deviation and $t_{n-1,1-\alpha/2}$ denotes the $100\left(1-\alpha/2\right)$th percentile of the t distribution with $n-1$ degrees of freedom. The value of $t_{n-1,1-\alpha/2}$ is always greater than $z_{1-\alpha/2}$; the values are close for large n, but t may be considerably larger than z for very small samples.

The required sample size would need to be larger than given by (1.30) simply to reflect the fact that $t_{n-1,1-\alpha/2} > z_{1-\alpha/2}$. In addition, the value of the standard deviation estimated from the sample will differ from the true standard deviation. The observed value of s may be either smaller or larger than the true value of the standard deviation, σ, and it can be expected to be larger than σ in about half of samples. So, even for large n, the observed confidence interval width will be greater than the specified ω in about half of planned studies.

To ensure that the observed confidence width will be shorter than ω more than half the time, we must take the distribution of s into account in the sample size computations. To solve for the required sample size for a confidence interval whose width has a specified probability, $1-\gamma$, of being narrower than ω requires the use of sample size software since an iterative solution based on the F and χ^2 distributions must be used (Kupper and Hafner 1989).

Returning to Example 10, specifying in nQuery Advisor that the observed confidence interval width needs to be shorter than 0.1 with a probability of 50% $(1 - \gamma = 0.5)$ yields a required sample size of 280, only slightly larger than that given by (1.30). However, to increase the likelihood that the observed confidence interval width will be shorter than ω from 50% to 90% would require an increase in sample size from 280 to 309 (see Table 1.9).

Table 1.10 shows the required sample sizes for two-sided 95% confidence intervals to have specified widths (expressed in terms of ω/σ).

Table 1.9. Confidence interval for mean based on t (with coverage probability)

	1	2
Confidence level, $1-\alpha$	0.950	0.950
1 or 2 sided interval?	2	2
Coverage probability, $1-\gamma$	0.500	0.900
Standard deviation, σ	0.850	0.850
Distance from mean to limit, ω	0.100	0.100
n	280	309

Table 1.10. Sample size for 95% two-sided confidence interval for a single mean to have width less than or equal to ω with probability

	$100\,(1-\gamma)$	
ω/σ	50%	90%
0.05	1539	1609
0.10	386	421
0.20	98	116
0.30	45	56
0.50	18	24

Note that nQuery Advisor has been used to compute the sample sizes displayed in all the rest of the tables in this chapter.

1.4.2 Hypothesis Testing for a Single Population Mean

Suppose we would like to test the hypothesis

$$H_0 : \mu = \mu_0$$

versus the alternative hypothesis

$$H_\mathrm{a} : \mu > \mu_0$$

and we would like to fix the level of the Type I error to equal α and the Type II error to equal β. That is, we want the power of the test to equal $1-\beta$. We denote the actual value of the population mean under the alternative hypothesis as μ_a. Following the same development as for hypothesis testing about the population proportion (with the additional assumption that the variance of $\bar{x}$ is equal to σ^2/n under both H_0 and H_a), the necessary sample size for this hypothesis testing situation is given by:

$$n = \frac{\sigma^2 \left[z_{1-\alpha} + z_{1-\beta}\right]^2}{\left[\mu_0 - \mu_\mathrm{a}\right]^2} \,. \tag{1.32}$$

Alternatively, defining the effect size as

$$\delta = \frac{\mu_0 - \mu_\mathrm{a}}{\sigma} \,, \tag{1.33}$$

we have

$$n = \frac{\left[z_{1-\alpha} + z_{1-\beta}\right]^2}{\delta^2} \,. \tag{1.34}$$

Example 11 . Pre and post studies with placebo in a variety of studies indicated that the standard deviation of blood pressure change was about 6 mm Hg and that the mean reduction in the placebo group was typically close to 5 mm Hg. To make a preliminary estimate of the value of a new intervention designed to lower blood pressure it was planned to enroll subjects and test the null hypothesis that mean reduction was 5 mm Hg. The new intervention would be of interest if the mean reduction was 10 or greater. How large a sample would be necessary to test, at the 5% level of significance with a power of 90%, whether the average blood pressure reduction is 5 mm Hg versus the alternative that the reduction is 10 mm Hg when it is assumed that the standard deviation is 6 mm Hg? ♦

Using (1.32) we have

$$n = \frac{6^2(1.645 + 1.282)^2}{(10-5)^2} = 12.33 \,.$$

Therefore, a sample of 13 patients with high blood pressure would be required.

A similar approach is followed when the alternative is two-sided. That is, when we wish to test

$$H_0 : \mu = \mu_0$$

versus

$$H_\mathrm{a} : \mu \neq \mu_0 \ .$$

In this situation, the null hypothesis is rejected if $\bar{x}$ is too large or too small. We assign area $\alpha/2$ to each tail of the sampling distribution under H_0. The only adjustment to (1.32) is that $z_{1-\alpha/2}$ is used in place of $z_{1-\alpha}$ resulting in

$$n = \frac{\sigma^2 \left[z_{1-\alpha/2} + z_{1-\beta}\right]^2}{\left[\mu_0 - \mu_\mathrm{a}\right]^2} \ . \tag{1.35}$$

Returning to Example 11, a two-sided test could be used to test the hypothesis that the average reduction in blood pressure is 5 mm Hg versus the alternative that the average reduction in blood pressure has increased, and that a reduction of 10 mm Hg would be considered important. Using (1.35) with $z_{1-\alpha/2} = 1.960$, $z_{1-\beta} = 1.282$ and $\sigma = 6$,

$$n = \frac{6^2(1.960 + 1.282)^2}{(10 - 5)^2} = 15.1 \ .$$

Thus, 16 patients would be required for the sample if the alternative were two-sided.

Since usually the true standard deviation is unknown, a more accurate solution for the necessary sample size would require use of sample size software (computations are based on the central and non-central t distributions). Unlike the situation for confidence intervals, the normal approximation formula works well for computing sample size for a test; its accuracy can be improved by adding the correction factor

$$\frac{z^2_{1-\alpha/2}}{2} \tag{1.36}$$

before rounding up. For Example 11 this would lead to a sample size estimate of 18 (which agrees with the result given by nQuery Advisor).

Table 1.11 presents the sample sizes necessary for 80% or 90% power for two-sided 5% level tests for specified effect sizes, δ.

Table 1.11. Sample size for two-sided 5% level t test to detect effect size $\delta = \mu_1 - \mu_2/\sigma$

δ	80% power	90% power
0.2	199	265
0.4	52	68
0.6	24	32
0.8	15	19
1.0	10	13
1.2	8	10

1.5 Comparison of Two Independent Means

1.5.1 Confidence Intervals for the Difference Between Two Means

The difference between two population means is represented by a new parameter, $\mu_1 - \mu_2$. An estimate of this parameter is given by the difference in the sample means, $\bar{x}_1 - \bar{x}_2$. The mean of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is

$$E\left(\bar{x}_1 - \bar{x}_2\right) = \mu_1 - \mu_2$$

and the variance of this distribution when the two samples are independent is

$$\text{Var}\,\left(\bar{x}_1 - \bar{x}_2\right) = \text{Var}\,\left(\bar{x}_1\right) + \text{Var}\,\left(\bar{x}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\,,$$

where n_1 and n_2 are the sample sizes in the two groups.

In order for the distribution of the difference in sample means, $\bar{x}_1 - \bar{x}_2$ to have a t distribution, we must assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. When the variances are equal and both sample sizes are equal to n, the formula for the variance of the difference can be simplified to

$$\text{Var}\,\left(\bar{x}_1 - \bar{x}_2\right) = \frac{2\sigma^2}{n}\,.$$

The value σ^2 is an unknown population parameter, which can be estimated from sample data by pooling the individual sample variances, s_1^2 and s_2^2 to form the *pooled variance*, s_p^2, where, in the general case,

$$s_\text{p}^2 = \frac{\left(n_1 - 1\right) s_1^2 + \left(n_2 - 1\right) s_2^2}{\left(n_1 - 1\right) + \left(n_2 - 1\right)}\,.$$

Example 12 . Nutritionists wish to estimate the difference in caloric intake at lunch between children in a school offering a hot school lunch program and children in a school that does not. From other nutrition studies, they estimate that the standard deviation in caloric intake among elementary school children is 75 calories, and they wish to make their estimate to within 20 calories of the true difference with 95% confidence. ♦

Using the normal approximation, the two-sided $100\left(1 - \alpha/2\right)$ % confidence interval for the true mean, $\mu_1 - \mu_2$, is given by

$$\bar{x}_1 - \bar{x}_2 \pm z_{1-\alpha/2} 2\sigma/\sqrt{n}\,. \tag{1.37}$$

So the sample size in each group required to obtain a confidence interval of width ω is

$$n = \frac{z^2_{1-\alpha/2} 2\sigma^2}{\omega^2} \; . \tag{1.38}$$

For Example 12,

$$n = \frac{(1.96)^2 (2)(75)^2}{(20)^2} = 108.05 \; .$$

Thus, a sample size of 109 children from each school should be selected.

We note, however, that the actual confidence interval for the difference in sample means would be given by

$$\bar{x}_1 - \bar{x}_2 \pm t_{2n-2,1-\alpha/2} s_{\mathrm{p}} \sqrt{2} / \sqrt{n} \; , \tag{1.39}$$

where s_{p} is the observed pooled standard deviation and $t_{2n-2,1-\alpha/2}$ denotes the $100\left(1-\alpha/2\right)$ percentile of the t distribution with $2(n-1)$ degrees of freedom. So, as explained in the section on confidence intervals for a single mean, to solve for the required sample size for a confidence interval whose width has a specified probability, $1-\gamma$, of being narrower than ω requires the use of sample size software.

For Example 12, we show in Table 1.12 (pasted from nQuery Advisor) that a sample of 109 per group provides a 50% probability that the observed 95% confidence interval will have half-width less than 20, while to have a 90% probability that the confidence interval half-width will be less than 20 would require a sample of 123 children per school.

Table 1.12. Confidence interval for difference of two means (coverage probability) (equal *n*'s)

	1	2
Confidence level, $1-\alpha$	0.950	0.950
1 or 2 sided interval?	2	2
Coverage probability, $1-\gamma$	0.500	0.900
Common standard deviation, σ	75.000	75.000
Distance from difference to limit, ω	20.000	20.000
n per group	109	123

Table 1.13. Sample size per group for 95% two-sided confidence interval for the difference in means to have width less than or equal to $\pm\omega$ with probability $\left(1-\gamma\right)$

	$100\left(1-\gamma\right)$	
ω/σ	50%	90%
0.05	3075	3145
0.10	770	805
0.20	193	211
0.30	87	98
0.50	36	39

Table 1.13 presents the sample sizes in each group required so that the two-sided 95% confidence interval for the difference in two independent means will be no wider than $\pm\omega$ with probability $(1-\gamma)$.

1.5.2 Testing the Difference Between Two Means (Two-Sample *t* Test)

The two-sample *t* test is used to test hypotheses about the population means in two independent groups of subjects. It is based on the assumptions that the underlying population distributions have equal standard deviations, and that the population distributions are Gaussian (normal) in shape or that the sample sizes in each group are large. (In most cases, the distribution of the sample mean will be approximately Gaussian for sample sizes greater than 30.)

We consider tests of the null hypothesis:

$$H_0 : \mu_1 = \mu_2 \quad \text{or}$$

$$H_0 : \mu_1 - \mu_2 = 0$$

versus either

$$H_{\mathrm{a}} : \mu_1 \neq \mu_2 \quad \text{for a two-sided test, or}$$

$$H'_{\mathrm{a}} : \mu_1 > \mu_2 \quad \text{or} \quad H''_{\mathrm{a}} : \mu_1 < \mu_2 \quad \text{for one-sided tests .}$$

To avoid repetitions of formulas with minor changes, we write formulas only in terms of a two-sided test.

The sample size required in each group, to achieve a power of $1-\beta$ is

$$n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_1 - \mu_2)^2} . \tag{1.40}$$

Setting

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} , \tag{1.41}$$

where δ is the effect size, we have a simpler version

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2} . \tag{1.42}$$

To improve the approximation, the correction factor in (1.43) may be added to (1.42) before rounding up.

$$\frac{z^2_{1-\alpha/2}}{4} . \tag{1.43}$$

Example 13. A two-group, randomized, parallel, double-blind study is planned in elderly females after hip fracture. Patients will be studied for two weeks; each patient will be randomly assigned to receive either new drug or placebo three times per week. The sample sizes in the two groups will be equal. Plans call for a 5% level two-sided t test. The outcome variable will be change in hematocrit level during the study. Prior pilot data from several studies suggests that the standard deviation for change will be about 2.0% and it would be of interest to detect a difference of 2.2% in the changes observed in placebo and treated groups. What sample size in each group would be required to achieve a power of 90%? ♦

For Example 13, the effect size is 2.2/2 = 1.1. Using (1.42) we find

$$n = \frac{2(1.96 + 1.28)^2}{(1.1)^2} = 17.4\ .$$

Adding the correction factor of 0.96 and rounding up, we have a required sample size of 19 per group, which is the solution given using nQuery Advisor (computations are based on iterative methods and the central and non-central t, see Dixon and Massey 1983 or O'Brien and Muller 1983).

Table 1.14 shows the sample size needed in each group for a two-sided 5% level t test to achieve 80% or 90% power for the specified alternative, δ.

Table 1.14. Sample size in each group for two-sided 5% level t test to have specified power

δ	80% power	90% power
0.2	394	527
0.4	100	133
0.6	45	60
0.8	26	34
1.0	17	23
1.2	12	16

1.5.3 Additional Considerations and References

Good introductions to sample size computations for tests and confidence intervals for a single mean or for comparing two independent means can be found in Dixon and Massey (1983), O'Brien and Muller (1983), Lemeshow et al. (1990), Lachin (1981), and Rosner (2000). Books containing sample size tables are available (e.g. Machin and Campbell 1987; Machin et al. 1997). Commercially available sample size software such as nQuery Advisor Release 6 (Elashoff 2005) can be used to compute sample size (or width) for confidence intervals and sample size or power for hypothesis tests for means for either the single group or two group designs, as well as for a wide variety of other sample size problems.

When plans call for the sample sizes in the two groups to be unequal, the formulas for sample size and power must incorporate the expected ratio of the sample sizes, see references above. For the two-sample t test, for any given total sample size, N, power will be highest when both groups have the same sample size. For this reason we generally prefer to plan equal sample sizes for a two-group study. However, sometimes investigators wish to plan a study with unequal n's; perhaps one type of subject is easier to accrue, or perhaps the investigator wants to maximize the number of subjects receiving the presumably superior treatment, or to accumulate extra safety information for the new treatment.

When the standard deviations in the two groups are markedly unequal, the usual t test with pooled variances is no longer the appropriate test. In many situations, the standard deviations show a patterned lack of homogeneity in which groups with higher means have higher standard deviations. In such cases, it is frequently advisable that sample size predictions (and later analysis) should be done on a transformed version of the variable. If the relationship between variance and mean is linear, this suggests using the square root of the variable. Such a transformation is likely to be desirable if the data represent counts or areas (note that the variable cannot be less than zero). If the relationship between standard deviation and mean is linear, this suggests using the log of the variable. This transformation is likely to be desirable for biological measures like viral load, triglyceride level, or variables ranging over several orders of magnitude (note that the variable cannot be negative or zero). If transformation does not seem to provide a solution to the problem of inequality of variances, it is possible that comparison of means is no longer the most appropriate method of analysis to address the question of interest. Assuming that transformation is not useful and comparison of means using a two-sample t test is still deemed appropriate, a modification of the t test may be planned; see for example, Moser et al. (1989) and sample size tables for the Satterthwaite t in nQuery Advisor.

If non-normality is an issue, planning a large study or considering transformations as above may be helpful; another possibility is to plan to use a non-parametric procedure instead, such as the two-sample Mann-Whitney/Wilcoxon rank test. For a description of this test, see Rosner (2000), and for methods to determine sample size and power see Hettmansperger (1984), Noether (1987), or sample size tables in nQuery Advisor.

Note that the sample size methods for comparisons of two independent means discussed above do not apply to correlation/agreement/repeated measures (or pair-matched case-control) studies in which N subjects are recruited and each subject is measured by two different raters, or is studied under two different treatments in a cross-over design. These designs cannot be analyzed using the methods described for independent means but must be analyzed using the paired t test or a repeated measures analysis of variance; see Rosner (2000) for information on the paired t test, and Muller and Barton (1989) or sample size tables in nQuery Advisor for information about sample size and power for repeated measures tests.

Logistic Regression Models

1.6

In prior sections of this chapter, we discussed sample size problems for estimation or testing of a proportion in one or two groups. In this section, we consider study designs in which it is planned to evaluate several predictor variables for a binary outcome variable. Specifically we consider studies in which we plan to fit a logistic regression model. Readers needing an introduction to the logistic regression model and test procedures should consult Hosmer and Lemeshow (2000).

In our experience there are two sample size questions, prospective and retrospective. The prospective question is: How many subjects do I need to observe to test the significance of a specific predictor variable or set of variables? The retrospective question is: Do I have enough data to fit this model? In this section we consider methods for choosing a sample size first and then discuss the importance of having an adequate number of events per covariate.

With respect to planning sample size for logistic regression, we distinguish two situations: (1) only a single covariate is of interest, (2) the addition of one covariate to a model already containing k covariates is of interest. In addition, we must distinguish whether the covariate of interest is dichotomous or continuous.

The basic sample size question is as follows: What sample size does one need to test the null hypothesis that a particular slope coefficient, say for covariate 1, is equal to zero versus the alternative that it is equal to some specified value.

Single Dichotomous Covariate

1.6.1

If the logistic regression model is to contain a single dichotomous covariate, then one may use conventional sample size formulas based on testing for the equality of two proportions. Hsieh et al. (1998) recommend using the following method to obtain sample sizes for logistic regression with a dichotomous covariate. (Although Whitemore 1981 provides a sample size formula for a logistic regression model containing a single dichotomous covariate, this formula, based on the sampling distribution of the log of the odds ratio, was derived under the assumption that the logistic probabilities are small and may be less accurate than the method we outline.)

Let the covariate X define two groups; group 1 contains those subjects for which $x = 0$ and the probability that the outcome of interest $y = 1$ for the subjects in this group is π_1, while group 2 contains those subjects for which $x = 1$ and the probability that $y = 1$ for these subjects is π_2.

Example 14 . Suppose that about 1% of the population is expected to have a particular adverse reaction to a certain drug used to treat a severe illness. It is thought that those with a specific pre-existing condition (expected to be about 20% of the population) will be much more likely to have such a reaction; it will be important to detect an odds ratio of two for the likelihood of a reaction in this group using a 5% two-sided likelihood ratio test. ♦

Table 1.15. Two group χ^2 test of equal proportions (odds ratio = 1) (unequal n's)

Test significance level, α	0.050	0.050	0.050
1 or 2 sided test?	2	2	2
No condition proportion, π_1	0.010	0.010	0.010
Pre-existing proportion, π_2	0.020	0.029	0.039
Odds ratio, $\psi = \pi_2(1-\pi_1)/[\pi_1(1-\pi_2)]$	2.000	3.000	4.000
Power (%)	90	90	90
n_1	7620	2468	1345
n_2	1905	617	337
Ratio: n_2/n_1	0.250	0.250	0.250
$N = n_1 + n_2$	9525	3085	1681

To compute the required sample size for Example 14 by hand would require using a modification of (1.27) for comparison of two proportions with unequal sample sizes, see references given in that section. Table 1.15 shows the table of results pasted from nQuery Advisor. (In this table, the symbol ψ is used to denote the odds ratio.) Defining group 1 as those without the pre-existing condition and group 2 as those with, the ratio of the sample size in group 2 to the sample size in group 1 will be 20/80 = 0.25. Using $\pi_1 = 0.01$ for group 1 (no pre-existing condition), and $OR = 2$, we find $\pi_2 = 2(0.01)/\big(2(0.01) + 0.99\big) = 0.02$. Table 1.15 shows that to detect an odds ratio of 2 with 90% power for this example would require a sample size of 9525. Consequently, the investigator may be interested in looking at the sample sizes required to detect odds ratios of 3 or of 4 (3085 and 1681 respectively).

1.6.2 Single Continuous Covariate

If the single covariate we plan to include in the model is continuous, approximate formulas for this setting have been derived by Hsieh (1989) and implemented in sample size software packages such as nQuery Advisor. However, Hsieh et al. (1998) demonstrate that this approximate formula gives larger than required sample sizes and recommend using the following method to obtain sample sizes for logistic regression with a continuous covariate.

Let the response Y define two groups; group 1 contains cases in which $Y = 1$ with $N\pi_1$ cases expected, while group 2 contains cases in which $Y = 0$ with $N\left(1-\pi_1\right)$ cases expected. The ratio of the expected sample size in group 2 to the expected sample size in group 1, r, is $\left(1-\pi_1\right)/\pi_1$. The natural log of the odds ratio, the coefficient β of the covariate, x, is equal to the difference between the mean of the covariate in group 1 and the mean of the covariate in group 2 divided by the within-group standard deviation of x (denote this by δ). Therefore, a sample size formula or table for the two group t test with unequal n's can be used to estimate sample size for logistic regression with one continuous covariate.

Example 15 . Patients with blocked or narrowed coronary arteries may undergo interventions designed to increase blood flow. Typically, about 30% of patients followed for a year will have renewed blockage, called "restenosis", of the artery. A study is to be planned to use logistic regression to assess factors related to the likelihood of restenosis. One such factor is serum cholesterol level. Based on the results of a large screening trial, mean serum cholesterol in middle-aged males is about 210 mg/dL; one standard deviation above the mean (which corresponds to about the 85th percentile) is approximately 250 mg/dL. In the screening study, the odds ratio for the six-year death rate for these two cholesterol levels was about 1.5. The study should be large enough to detect an effect of serum cholesterol on arterial restenosis of a size similar to that seen for death rate. We plan to conduct the test of the predictive effect of cholesterol level on the probability of restenosis using a 5% two-sided test and want to have 90% power to detect an odds ratio of 1.5 for values of cholesterol of 250 mg/dL versus 210 mg/dL. We set the effect size, $\delta = \left(\mu_1 - \mu_2\right)/\sigma = 0.405$, which is the value of the natural log of the odds ratio, 1.5. The ratio of sample sizes expected to be in the no-restenosis versus the restenosis groups, r, equals 0.7/0.3 = 2.333. ♦

The required sample size could be computed using a version of (1.42) modified for unequal sample sizes, see references in the preceding section. In Table 1.16 we show the table of results pasted from the software nQuery Advisor.

Table 1.16. Two group t-test of equal means (unequal n's)

Test significance level, α	0.050
1 or 2 sided test?	2
Effect size, $\delta = \left\|\mu_1 - \mu_2\right\|/\sigma$	0.405
Power (%)	90
n_1	93
n_2	217
Ratio: n_2/n_1	2.333
$N = n_1 + n_2$	310

To obtain a power of 90% to detect an odds ratio of 1.5 using the covariate cholesterol to predict restenosis at one-year, we find that a total sample size of 310 is required.

1.6.3 Adjusting Sample Size for Inclusion of k Prior Covariates (Variance Inflation Factor)

It is rare in practice to have final inferences based on a univariate logistic regression model. However, the only sample size results currently available for the multivariable situation are based on very specific assumptions about the distributions of the covariates. We can however, use a "variance inflation factor" to adjust the sample size results obtained for a single covariate for the situation in which k covariates have already been added to the model before the new covariate is considered.

The sample size, N_k, required to test for the significance of a covariate after inclusion of k prior covariates in the model, is given by

$$N_k = N\left(\frac{1}{1-\varrho^2}\right) , \tag{1.44}$$

where the factor $1/\left(1-\varrho^2\right)$ is called the "variance inflation factor",

$$VIF = \left(\frac{1}{1-\varrho^2}\right) , \tag{1.45}$$

and ϱ^2 is the squared multiple correlation of the covariate of interest with the covariates already included in the model. This can be estimated using any multiple regression software.

For Example 14, the total sample size was computed as $N = 1681$ for testing the significance of one dichotomous covariate. Now assume that four patient demographic variables will be entered into the logistic regression model prior to testing the covariate indicating presence or absence of the pre-existing condition (x_1 say), and that these demographic variables have a squared multiple correlation with x_1 of 0.2. Then a total sample size of at least 2100 patients would be required,

$$N_4 = 1681\left(\frac{1}{1-0.2}\right) = 2101 .$$

In Example 15 if two other covariates with a squared multiple correlation with cholesterol of 0.15 are to be entered into the logistic regression first, multiply the sample size obtained for a single covariate by the variance inflation factor, (1.44), $1/\left(1-\varrho^2\right) = 1.18$, to increase the required sample size to 365.

1.6.4 Assessing the Adequacy of Data Already Collected

So far we have discussed planning what sample size should be obtained to fit specific logistic regression models. A second consideration, and one relevant to any model being fit, is the issue of what is the maximum number of covariates it is reasonable to enter into the model and still obtain reliable estimates of the regression coefficients and avoid excessive shrinkage when the model is assessed for new cases. An ad hoc rule of thumb is to require that there be 10 "events" per variable included in the model. Here the "event" of relevance is the least frequent of the outcomes. For example, suppose the study discussed in Example 15 was planned with 365 cases. Further suppose that complete one-year follow-up was only obtained for 351 cases of which 81 had restenosis at one year. There are 81 cases with restenosis and 270 without, so the least frequent "event" is restenosis. Based on these 81 cases, only 8 variables should be fit; this means that no more than 8 covariates (or covariates plus covariate interaction terms) should be entered into the model.

This rule of thumb was evaluated and found to be reasonable by Peduzzi et al. (1996) using only discrete covariates. However, as is the case with any overly simple solution to a complex problem, the rule of 10 should only be used as a guideline and a final determination must consider the context of the total problem. This includes the actual number of events, the total sample size and, most importantly, the mix of discrete, continuous and interaction terms in the model. The "ten events per parameter" rule may work well for continuous covariates and discrete covariates with a balanced distribution. However, its applicability is less clear in settings where the distributions of discrete covariates are weighted heavily to one value.

Practical Issues in Sample Size Choice

In earlier sections, we outlined formulas for sample size computation for estimation and testing in simple designs for proportions and for means. We have shown only formulas to compute sample size from specifications of confidence interval width or desired power, but it is also possible to compute the confidence interval width or power which would be obtainable with a specified sample size. Sample size methods exist for many more complex designs and for other parameters. Software such as nQuery Advisor (Elashoff 2002) can be helpful.

For complex study designs or complex statistical methods, however, there may be no easily applied formulas or available software solutions. In such cases, sample size choices may be based on simplifications of the design or statistical methods (as we illustrated in the section on logistic regression), or in some cases a simulation study may be warranted.

For studies involving complex survey designs, sample size computations might be based on one of several approaches: (1) regarding the cluster itself as the study "subject" and using intraclass correlation values to estimate the appropriate variance to use in making computations, (2) multiplying sample sizes for a simpler design by a computed "design effect" (2 may be a sensible ad hoc choice), or (3) using simulation methods.

Although study sample sizes are usually chosen to assure desired precision or power for the primary outcome variable, investigators may also need to investigate whether that sample size choice will be adequate for evaluations of secondary outcomes, or for analyses of pre-defined subsets.

Sample size values obtained from formulas or software will generally need to be inflated to allow for expected dropout or loss to followup of study subjects or other sources of missing data (cf. Chap. II.6 of this handbook). It is important to remember however, that subjects who drop out may not be similar to those remaining in the study. This consideration may affect the parameter values which should be used for sample size computations; and even analyses using missing data techniques may not remove biases due to dropout.

Another issue of great concern to epidemiologists is that exposure or response may be misclassified. Such misclassification might have a dramatic impact on the actual power of a planned study unless sample sizes are computed based on modeling the expected type and extent of misclassification using simulation methods.

For brevity, our examples used only one set of parameter values to compute required sample sizes. In practice, investigators need to keep in mind that the estimated parameter values used in computations are only estimates and perhaps not very accurate ones. It is a good idea to compute necessary sample size for several different sets of parameter choices to evaluate sample size sensitivity to varying realistic possibilities for the true parameter values. Tables and plots can be helpful in these evaluations.

Finally, sample size justification statements in protocols, grant proposals, and manuscripts need to be complete. Details of the outcome variable, the study design, the planned analysis method, confidence level or power, one or two-sided, and all the relevant distributional parameters (proportions, means, standard deviations) need to be included in the statement. For Example 13 a minimal sample size justification might read as follows: *A sample size of* 19 *in each group will have* 90% *power to detect a difference in means of* 2.2 *(the difference between an active drug mean change in hematocrit of* 2.2% *and a placebo mean change of* 0.0*) assuming that the common standard deviation is* 2.0 *and using a two group* t test *with a* 0.05 *two-sided significance level. The planned enrollment will be* 25 *subjects* per group (50 total) to allow for 20% dropout. It is also desirable to provide information about sample size for other parameter choices and details about how these parameter values were selected, including references to previous studies which were consulted in selecting the values.

1.8 Conclusions

An important part of planning any research study is to assess what sample size is needed to assure that meaningful conclusions can be drawn about the primary outcome. To do this, the investigator must detail the study design, define the primary outcome variable, choose an analysis method, and specify desired or expected results of the study. Then formulas, tables, and sample size software of the sort outlined in this chapter can assist with computations. The most essential part of the process, though, is to make a thorough investigation of other information and research results concerning the outcome variable to support reasonable specification of hypothesized values for use in making computations. Beginning investigators often protest: "But this study has never been done before; how do I know what the results will be?" In most cases, however much information about rates, means, and standard deviations can be gleaned from other contexts and used to infer what kinds of outcomes would be important to detect or likely to occur. Sample size computations are not just a pro forma requirement from funding agencies but provide the basis for deciding whether a planned study is likely to be worth the expense.

References

Chernick MR, Liu CY (2002) The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods. The American Statistician 56:149–155

Dixon WJ, Massey FJ (1983) Introduction to statistical analysis. 4th edn. McGraw-Hill, New York

Elashoff JD (2005) nQuery Advisor® Release 6. Statistical Solutions, Ireland

Fleiss JL (1981) Statistical methods for rates and proportions. 2nd edn. Wiley, New York

Fleiss JL, Tytun A, Ury SH (1980) A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 36:343–346

Hettmansperger TP (1984) Statistical inference based on ranks. Wiley, New York

Hosmer DW, Lemeshow S (2000) Applied logistic regression. 2nd edn. Wiley, New York

Hsieh FY (1989) Sample size tables for logistic regression. Statistics in Medicine 8:795–802

Hsieh FY, Bloch DA, Larsen MD (1998) A simple method of sample size calculation for linear and logistic regression. Statistics in Medicine 17:1623–1634

Korn EL (1986) Sample size tables for bounding small proportions. Biometrics 42:213–216

Kupper LL, Hafner KB (1989) How appropriate are popular sample size formulas? The American Statistician 43:101–105

Lachin JM (1981) Introduction to sample size determination and power analysis for clinical trials. Controlled Clinical Trials 2:93–113

Lachin JM (1992) Power and sample size evaluation for the McNemar test with application to matched case-control studies. Statistics in Medicine 11:1239–1251

Lemeshow S, Hosmer DW, Klar J, Lwanga SK (1990) Adequacy of sample size in health studies. Wiley, Chichester

Levy PS, Lemeshow S (1999) Sampling of populations: methods and applications, 3rd edn. Wiley, New York

Louis TA (1981) Confidence intervals for a binomial parameter after observing no successes. The American Statistician 35:154

Machin D, Campbell MJ (1987) Statistical tables for design of clinical trials. Blackwell Scientific Publications, Oxford

Machin D, Campbell M, Fayers P, Pinol A (1997) Sample size tables for clinical studies. 2nd edn. Malden and Carlton: Blackwell Science, London

Moser BK, Stevens GR, Watts CL (1989) The two-sample *t* test versus Satterthwaite's approximate F test. Commun. Statist.-Theory Meth. 18:3963–3975

Muller KE, Barton CN (1989) Approximate power for repeated-measures ANOVA lacking sphericity. Journal of the American Statistical Association 84:549–555

Noether GE (1987) Sample size determination for some common nonparametric tests. Journal of the American Statistical Association 82:645–647

O'Brien RG, Muller KE (1983) Applied analysis of variance in behavioral science. Marcel Dekker, New York, pp 297–344

Peduzzi PN, Concato J, Kemper E, Holford TR, Feinstein A (1996) A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology 99:1373–1379

Rosner B (2000) Fundamentals of Biostatistics. 5th edn. Duxbury Press, Boston

Whitemore AS (1981) Sample size for logistic regression with small response probability. Journal of the American Statistical Association 76:27–32