CHAPTER 2

Methods of Determining Sample Sizes

Sample size and power determinations are an important and necessary component of research grant proposals, including proposals involving clinical trials submitted to the National Institutes of Health (NIH). Biostatisticians often assist in the statistical part of grant proposals and are often a member of the research team. The guidelines for researchers provided by the Quantitative Methods Core of the Department of Quantitative Health Sciences in the University of Massachusetts Medical School may be of interest (`http://www.umassmed.edu/uploadedFiles/QHS/QMCGuidelinesforGrantApplications2.pdf`).

Before embarking on a study of sample size determination and power, it is worth noting that there are two relatively recent papers that rattle conventions and that should be read by anyone faced with the task of determining sample size, whether it be infrequently as in writing grant proposals, or for more frequent applications as might be involved with a company's operations. Those two papers are Bacchetti, Deeks, and McCune (2011) and Bacchetti (2010). Quoting from the latter: "Inaccuracy of sample size calculations is not only theoretically inevitable (Kraemer, Mintz, Noda, Tinklenberg, and Yesavage, 2006; Matthews, 1995) but also empirically verified." Bacchetti (2010) went on to relate that Vickers (2003) found in a study of assumed values for standard deviations for randomized clinical trials published in four leading general medical journals that about one-fourth had more than a fivefold inaccuracy in sample size and more than half had more than a twofold inaccuracy. Bacchetti (2010) also stated that it is difficult to estimate a standard deviation accurately unless there is such a large amount of preliminary data available that the planned study is unnecessary! The author debunks the idea of a minimum acceptable power of .80 and instead advocates Value of Information methods, for which sample size is determined to maximize the expected value of the information produced minus the total cost of the study.

Sample Size Determination and Power, First Edition. Thomas P. Ryan.

The author pointed out that it is easier to estimate cost than to estimate a standard deviation. This is a very important point since estimation of "other" parameters is necessary when determining sample size. A similar approach is advocated by Bacchetti, McCulloch, and Segal (2008), who stated that "we see no justification for ignoring costs." The response to Bacchetti (2010) by Schulz, Moher, and Altman (2010) may also be of interest, in addition to Bacchetti's reply. See also the determination of sample size when there is a cost constraint, as discussed by Guo, Chen, and Luh (2011)

The same general line of thinking is found in Bacchetti et al. (2011), who stated: "Studies of new ideas often must start small (sometimes even with an *n* of 1) because of cost and feasibility concerns, and recent statistical work shows that small sample sizes for such research can produce more projected scientific value per dollar spent than larger sample sizes." Glick (2011a) reviewed sample size and power formulas for cost-effectiveness analyses that have been given in the literature and Glick (201lb) explained the relationship between sample size and power and "maximum willingness to pay."

The methods espoused in these recently published papers, almost all of which have had Bacchetti as a co-author, may eventually lead to the use of new methods of determining sample size. That may take some time, however, so the emphasis in this book is on the conventional approach of determining sample size using power considerations, in addition to determining sample size for confidence intervals. There is also an emphasis, for simplicity and for space considerations, on determining a single sample size for a given problem, while recognizing the importance of the following quote from Sahu and Smith (2006): "A single reported sample size is not very informative on its own and its sensitivity with respect to many different assumptions should be investigated. In a practical situation this sensitivity needs to be explored and matched with the practical information that is available to decide the sample size."

Determining sample size does require overcoming certain obstacles. In particular, whether our objective is to determine a sample size so as to have a desired degree of power for a hypothesis test, to have a confidence interval with a specified width, or to estimate a parameter with a maximum error of estimation for a specified probability, it is necessary to specify values of the parameters that are involved in the distribution of the estimate of the parameter for which the hypothesis test, confidence interval, or point estimate is being constructed. Since the other parameters will also generally be unknown, some type of estimate must be provided for them. For example, in the outline given in the link provided in the first paragraph of this chapter, note B2(b): Continuous outcomes: need standard deviation. Of course, a sample estimate won't work (unless from a previous study) since the sample whose size is being determined hasn't been taken yet!

This naturally leads to consideration of Bayesian approaches, as pointed out by Adcock (1997), who discussed both frequentist and Bayesian approaches, as well as to sequential approaches, in general, as discussed by Betensky and Tierney (1997). A frequentist approach is the standard textbook approach to sample size determination and data analysis, with Bayesian approaches being more involved. Their one obvious commonality, however, is that both are used to provide a prior estimate of parameters other than the parameter that is being either tested in a hypothesis test or for which a confidence interval is to be constructed.

Inoue, Berry and Parmigiani (2005) pointed out that an increasing number of submissions for approval of drugs and medical devices, for example, are using Bayesian approaches. These authors also indicated, though, that "one of the problems in Bayesian submissions is that there is no standard method for calculating sample size from a Bayesian perspective." This coupled with possible difficulty in justifying a selected prior distribution for the parameters that must be specified may limit the number of studies in which Bayesian methods will be used. Joseph, du Berger, and Bélisle (1997) reviewed the use of a Bayesian approach to sample size determination and provided a good motivation for such an approach, with their focus being confidence interval width rather than hypothesis tests. Lindley (1997) also gave a Bayesian approach using a utility function, with Pezeshk, Nematollahi, Maroufy, and Gittins (2009) being related work. Earlier work on a Bayesian approach to sample size determination included Goldstein (1981), Adcock (1988), Weiss (1997), and Joseph and Bélisle (1997). See also Dendukuri, Rahme, Bélisle, and Joseph (2004) for Bayesian sample size determination with nonidentified models and Gustafson (2006), who considered sample size determination when model bias is modeled rather than ignored. Wang and Gelfand (2002) proposed a generic Bayesian approach that would be applicable to a wide range of models. A Bayesian sample size determination for equivalence testing and noninferiority testing was given by Wang and Stamey (2010). See also De Santis (2004, 2006), Pham-Gia and Bekker (2005), Stüger (2006), and Yin, Choudhary, Varghese, and Goodman (2008).

Although the use of Bayesian methods in sample size determination can thus be easily motivated, they are covered briefly in this chapter and in certain subsequent chapters because the methods are generally not available in the leading software for sample size determination. Readers who desire a more detailed and technical discussion of Bayesian sample size determination methods are referred to Chapter 13 of Chow, Shao, and Wang (2008).

Hare (2008) touched upon the problems that can be encountered with nonexperts who will say things like not to bother with a statistically rigorous approach to sample size determination, or that the sample size used should be "the industry standard," or that the sample size should be "where the numbers stabilize." None of these responses are valid, as Hare (2008) noted.

Similarly, ad hoc approaches can be found in fields in which methods are employed that lack mathematical rigor. For example, Ariola (2006, p. 140) stated: "These formula [sic] may either be Slovin's formula, Parten's formula, Ibe's formula, or the majority rule formula." None of these formulas are given in sampling books. On page 79, the author stated: "If there is ignorance of population, use Slovin's formula"

$$n = \frac{N}{\left(1 + Ne^2\right)}$$

with n = sample size, N = population size, and e = "desired margin of error." There are multiple problems with such a formula, which we might call a nonparametric sample size formula because there is no indication of a distributional assumption. While it is certainly true that there is no distributional assumption when nonparametric statistical methods are used, the size of a sample must certainly be a function of the population variability of whatever is being measured, and here there is no variability represented by the formula. If the numbers in a population differ very little, a small sample size could lead to a good estimate of the population mean, whereas a larger sample would be needed if the variability was not small. This point is made in every book on sampling.

Another obvious problem with Slovin's formula is that there is no indication of what the "margin of error" is for—a mean, a proportion, a parameter estimate? Although simple approaches will generally be preferred by practitioners, simple approaches that are flawed and not statistically justifiable should be eschewed.

One thing that should be kept in mind is that subjects may drop out of an experimental study, and/or some subjects may not comply with the stipulated conditions of the study. So the initial sample size and the "final" sample size may be quite different, and this should be taken into consideration in selecting the sample size for the beginning of a study. Furthermore, in various applications, such as medical applications, there will be sequential sample size selection because more than one phase will be involved, such as in clinical trials, and sequential sample size determination can also be used within a given phase. That is, sample size reestimation can occur as a study progresses. Shih (2001) provided an invited commentary on sample size reestimation, and the following quote (p. 515), in reference to clinical trials, should be noted: "Hence, one would very much like to utilize the information from the current trial at an interim stage to update the initial estimates and make adjustment of the sample size, if necessary, to ensure that the study's objective is accomplishable with adequate power."

2.1 INTERNAL PILOT STUDY VERSUS EXTERNAL PILOT STUDY

This leads to consideration of an internal pilot study versus an external pilot study. The former is part of the main study, whereas the latter is conducted for the purpose of obtaining necessary parameter estimates. For example, if a test of a population mean is being conducted, the sample size cannot be determined without having an estimate of the population standard deviation. A small (external) pilot study might be conducted before the main study for the express purpose of obtaining data that would be used to provide an estimate of any so-called nuisance parameter, such as the standard deviation. (A nuisance parameter is any parameter that is not the parameter of interest.) Wittes and Brittain (2008) cautioned against the use of external pilot studies for parameter studies, at least for clinical trials, however, as they stated, “external pilot studies are often very unrepresentative of the population to which they refer.” Instead, they recommended that internal pilot studies be used for parameter estimation. That is, at a certain point in the study/trial, the data that had been accumulated to that point would be used for parameter estimation, which would then enable the sample size to be recalculated, assuming that an initial estimate of sample size was used before the internal pilot study began.

The idea of using an internal pilot study for parameter estimation has also been discussed by many other authors, including Coffey and Muller (2001), Posch and Bauer (2000), Denne and Jennison (1999), and Browne (1995), as mentioned by Machin and Campbell (2005). Day (2000) reviewed previous work on internal pilot studies and addressed the question of the point at which to stop collecting data in an internal pilot study.

The general idea of an internal pilot study is to use an initial group of subjects, such as patients, with the sample size for the full study reestimated based on data from this initial group. The difference between this type of pilot study and a standard (external) pilot study is that with the latter the subjects are not part of the sample that is used in the regular study, whereas with an internal pilot study they are part of the sample and are thus “internal” to the study. This does have a consequence, however, as the Type I error rate is slightly inflated (Lancaster, Dodd, and Williamson, 2004) since the pilot study and the main study are not independent, as is tacitly assumed when the pilot study is used to estimate sample size for the main study. Wittes, Schabenberger, Zucker, Brittain, and Proschan (1999) found that the Type I error rate inflation is often negligible, however. Wittes and Brittain (1990) and Birkett and Day (1994) gave simulation results that quantified the extent to which the Type I error rate is inflated under certain conditions. Coffey and Muller (2001) gave analytical results that agreed with the simulation results of Wittes and Brittain (1990) and gave exact results that apply to any general linear univariate model with fixed predictors, such as regression models with fixed predictors, mentioned in Chapter 5, and ANOVA models of Chapter 6. Kieser and Friede (2000) explained that the t-test that results from use of a pilot study does not have a central t-distribution and derived the actual distribution, in addition to providing the expression for the actual Type I error and methods for controlling the Type I error level. Denne and Jennison (1999) stated that, when using an internal pilot study, the power of a t-test is not robust to misspecification of the variance and proposed a t-test for a two-treatment comparison that was based on Stein’s (1945) two-stage test.

Similarly, Lancaster et al. (2004) stated that in some situations patients involved in an external pilot study are later included in the main study, as a way of minimizing the number of people who must be recruited. This would also inflate the Type I error rate.

If the initial variance estimate that resulted in the initial sample size was at least equal to the sample variance from the internal pilot study, then there would be no adjustment to the initial sample size. If, however, the sample variance exceeded the initial variance estimate, there would be an appropriate adjustment in the sample size. The study would then continue using either the original sample size or a revised sample size obtained using the pilot study parameter estimates.

This obviously raises the question of the point at which the study should be stopped to obtain the parameter estimates. Wittes and Brittain (1990) do not address this issue. If the study is stopped too soon, there will be a large variance associated with the estimate of the population variance, so the latter might be poorly estimated, which in turn would produce a sample size that is either too large or too small. A conservative approach would be to construct a confidence interval for the population variance and use the upper limit of the interval in determining the sample size. While focusing on external pilot studies, Lancaster et al. (2004) cited Browne (1995) as recommending at least an 80% upper one-sided confidence limit, while also citing Browne (1995) as stating that a general rule of thumb is to use at least 30 patients. (That rule of thumb might be based on what is discussed in an introductory statistics book regarding large versus small samples.) Of course, these suggestions could also be applied to an internal pilot study, with the use of confidence limits for σ^2 also suggested by Dudewicz (1972), with this resulting in approximate confidence limits on power.

Whether or not this would be a reasonable approach would depend on the costs, both tangible and intangible, of using a larger-than-necessary sample size since the larger the variance estimate, the larger the sample size.

Some insight as to desirable sizes of internal pilot studies can be gleaned from the Pilot Study routine in Lenth's applet, which is mentioned briefly in Section 2.11 and is used in succeeding chapters. The applet routine essentially allows a user to see how the size of a pilot study will determine how well the sample size for the main study is determined. Specifically, the user enters the degrees of freedom for the "error term," with the latter being just σ for a test of one mean, specifies the percent by which the sample size is underestimated, and then sees as output an approximation of the probability that the percent underestimation is exceeded, with the probability being exact under certain conditions.

For example, if the degrees of freedom for error is 19 and the percent underestimation is 10, the applet gives a 41.69% chance that the percent underestimation will exceed 10, whereas this drops to 24.68% when the degrees of freedom is 100, and drops to only 9.44% when the degrees of freedom is 335. Of course, the latter might be prohibitively large for an external pilot study, but perhaps not for an internal pilot study, depending on the nature of the study. If the sample size percent underestimation was 15, a pilot study sample size of only 140 would be necessary for the percent risk to be (barely) less than 10, but such a percent underestimation could cause a study to be significantly underpowered.

Those numbers are obtained as follows. Consider the test of a population mean under the assumption of a normal distribution and an internal pilot study of n observations.

A well-known result from statistical theory is that $[(n-1)\,s^2]/\sigma^2 \sim \chi^2_{n-1}$, with "$\sim$" read "is distributed as" and χ^2_{n-1} denoting a chi-square random variable that has the chi-square distribution with $(n-1)$ degrees of freedom.The expected value (i.e., mean) of that random variable is $(n-1)$, which of course is what $[(n-1)\,s^2]/\sigma^2$ is equal to if $s^2 = \sigma^2$. Similarly, $s^2 \sim [\sigma^2/(n-1)]\chi^2_{n-1}$, so that $E(s^2) = \sigma^2$, taking the expected value of each side. If, for example, we knew that $s^2 = 0.9\sigma^2$, so that the use of s^2 to estimate σ^2 would, in expectation, underestimate the latter by 10%, then we need the value of χ^2_{n-1} such that

$$E\left[\left(\frac{\sigma^2}{n-1}\right)\chi^2_{n-1}\right] = 0.9\sigma^2$$

which means that we need $E\left(0.9\chi^2_{n-1}\right) = 0.9(n-1)$. Using the example with degrees of freedom = 100 from the ' preceding paragraph and still assuming that the percent underestimation is 10%, we need to obtain $P\left(\chi^2_{100} < 90\right)$. The use of statistical software shows that this probability is 0.2468, in agreement with the probability obtained using Lenth's applet. [*Note*: Here we are looking at the percent underestimation of σ^2, not σ, but it is the former that is most relevant because sample size formulas are a function of σ^2, as will be seen, for example, in Eq. (2.4) in Section 2.3.]

Table 2.1 shows the probability of underestimating the required sample size for selected values of n for the hypothesis test of a population mean, assuming a normal distribution. Such a table might be used to select the size of an internal pilot study.

These numbers are exact for the underestimation of the degrees of freedom for estimating σ and are thus approximations relative to the required sample size. The numbers show that it is highly desirable to have a reasonably large internal pilot study, but certainly not so large that the required sample size for the full study is less than the size of the internal pilot study.

Of course, this leads to the question of the extent to which an undersized full study results in the full study being underpowered. Power depends on other factors in addition to the sample size, specifically the significance level, the difference between the value of the mean under the null hypothesis and the value that is to be detected with the prescribed power if the null hypothesis is false, and σ. Therefore, a simple result cannot be given and further discussion of the effect on power is deferred to the examples that are given in later sections of this chapter. Some insight is given by Anderson (2006), however.

Table 2.1 Probability of Underestimating Required Sample Size[a]

Size of Internal Pilot Study	Percent by Which d.f. (σ) Is Underestimated	Probability of Exceeding Percentage
10	10(0.9)	.4759
	25(2.25)	.3369
	50(4.5)	.1245
	75(6.75)	.0131
25	10(2.4)	.3969
	25(6.0)	.1970
	50(12.0)	.0201
50	10(4.9)	.3279
	25(12.25)	.0985
	50(24.5)	.0013
100	10(9.9)	.2480
	25(24.75)	.0298
	50(49.5)	<.0001
200	10(19.9)	.1589
	25(49.75)	.0034
500	10(49.9)	.0532

[a] Note that since d.f. $= n - 1$, all of the percentages except the last two for size = 25 result in a degrees of freedom that is not an integer, which would correspond to the sample size for the full study being a noninteger value. No adjustment was made for that since Lenth's applet gives d.f. (σ).

2.2 EXAMPLES: FREQUENTIST AND BAYESIAN

A frequentist approach is contrasted with a Bayesian approach in Example 2.1, but Bayesian sample size determination techniques are not emphasized in this book, nor are they incorporated into the leading sample size determination software. Such software is available, however, as functions written for R and for S-Plus. See `http://www.medicine.mcgill.ca/epidemiology/joseph/Bayesian-Software-Bayesian-Sample-Size.html`.

Bayesian sample size determination is discussed further in this chapter in Section 2.1.1 and is also discussed, albeit somewhat briefly, in some of the subsequent chapters. Readers with an interest in Bayesian approaches to sample size determination are referred, in particular, to the references at the end of these chapters and to Chapter 13 of Chow, Shao, and Wang (2008). Also recommended is Inoue et al. (2005), who provided a general framework that allowed Bayesian approaches to sample size determination to be contrasted and compared with frequentist approaches.

We will start with a very simple, hypothetical example to illustrate the frequentist approach.

■ EXAMPLE 2.1

Let's assume that a manufacturing process yield is standardized in such a way that 100 has been the long-term value and this value has been acceptable. If intended improvements are made, however, there is a good chance that the yield will reach 101, and it is desired to determine the number of units of production to sample so that such an improvement can be detected, if it has occurred, with a probability of .80 (i.e., the power is to be .80, a value that is customarily used but will be critiqued in later sections).

That is, a hypothesis test will be performed such that the null hypothesis is $\mu = 100$ and the alternative hypothesis is $\mu > 100$.

What should be the sample size? We will assume an infinite population size for this example, as populations are frequently so large that they are practically countably infinite, and thus might as well be considered as infinite. Later we will look at how to proceed when populations are finite and the population size is known. We will also assume that the population can be adequately represented by a normal distribution and that σ, the standard deviation of the individual observations, is the same for both the distribution with the hypothesized mean and the distribution with the true mean. This is a reasonable assumption in the absence of any prior information to suggest otherwise, especially if the true mean does not differ greatly from the hypothesized mean.

The starting point is to look at the probability statement that leads to the sample size expression. If we use a significance level of .05, the statement is

$$P(\overline{X} > 100 + 1.645\sigma/\sqrt{n} | \mu = 101) = .80 \tag{2.1}$$

with $\sigma/\sqrt{n}$ being $\sigma_{\bar{x}}$ and $\bar{X}$ denoting the average of the individual observations, X, in a sample of size n. The value of $100 + 1.645\sigma/\sqrt{n}$ is the value for $\bar{X}$ such that values of $\bar{X}$ greater than this will result in the null hypothesis, $\mu = 100$, being rejected and the conclusion will be that $\mu > 100$. The constant 1.645 is determined by the .05 significance level and the fact that the test is one-sided, and is based on the assumption that either the individual observations have approximately a normal distribution (as stated for this example), or that the sample size, n, to be determined, will be large enough that the distribution of $\bar{X}$ will be approximately normal. The probability of the null hypothesis being rejected is .80 when $\mu = 101$ because the sample size will be determined so as to (approximately) produce that power if in fact $\mu = 101$. Note that we will conclude that $\mu > 100$ with probability .05 when $\mu = 100$ because we have selected .05 as the significance level. Note also that $1.645\sigma/\sqrt{n}$ goes to zero as $n \rightarrow \infty$, so if n was somewhat arbitrarily selected and made very large, eschewing sample size determination formula, the probability that $\bar{X}$ exceeds $100 + 1.645\sigma/\sqrt{n}$ could, depending on the value of σ, be virtually 1.0 not only for $\mu = 101$ but also for values of $\mu < 101$ that would not be of practical interest. Thus, the test would be too sensitive. This is one reason why a sample size should not be arbitrarily or haphazardly chosen, such as a sample size that uses all of the money that has been committed to a study.

A value for σ must be used unless the experimenter is willing to express in standard deviation units the shift that is to be detected with whatever probability is selected. The use of standard deviation units in sample size determination is not advocated here, however, and has also been criticized by Lenth (2001). The reason for this is perhaps obvious: standard deviation units enable the experimenter to avoid facing the question of what threshold parameter value defines practical significance in a given setting, plus the estimation of the standard deviation is avoided. Since the improvement of processes entails reducing variability, trying to avoid thinking about variability is not a good idea.

If a pilot study had been conducted, the sample standard deviation, s, could be used in estimating σ if the pilot study was of at least moderate size, although it should be kept in mind that s is a statistically biased estimator of σ. If the data were normally distributed (as assumed here), $E(s) = c_4\sigma$, with "$E(s)$" denoting the expected value of s and the value of c_4 is dependent on the sample size, being less than 1.0 for all sample sizes. For example, $c_4 = 0.9727$ for $n = 10$ and $c_4 = 0.9896$ for $n = 25$. Thus, for small samples it would be better to use s/c_4 to estimate σ, although it might be even better to use the upper limit of a confidence interval for σ, depending on the application. Tables of c_4 values are given in statistical quality control and improvement books, such as Ryan (2011). When $n \geq 10$, a reasonable approximation for c_4 is $c_4 = 1 - 1/4n - 7/32n^2$. There is not much point in using c_4 when n is much larger than 25, provided that there is evidence of at least approximate normality.

It is well known that $\text{Var}(s) = \sigma^2(1 - c_4^2)$, so a large value of σ^2 coupled with a small pilot study could result in sampling variability in s of such a magnitude that it cannot be ignored. Nevertheless, Sims, Elston, Harris, and Wanless (2007) stated: "In fields such as the physical and chemical sciences and engineering, detailed knowledge of the variance can often be obtained from a pilot study reasonably quickly and efficiently, and it is considered acceptable to regard the estimate as the true population variance. In some areas of biological research, such as clinical trials where variance information can be cumulated across trials with identical protocols,the assumption of known variance may also be valid." Since $\text{Var}(s) \to 0$ as $c_4 \to 1$, which occurs as $n \to \infty$, it is really a question of how much data is available for estimating σ before the major study is performed. At the other extreme, Sims et al. (2007) also stated: "In contrast to other fields of science, ecological pilot studies can be financially expensive and, where years act as a basic unit of replication, only accrue information very slowly." Thus, practitioners need to think about the field in which they are working in terms of the cost of pilot studies and the amount of data available for estimating σ. If in doubt, it would be best to *not* act as if σ is known and to incorporate sampling variability into the determination of power and sample size. Of course, this is even more important when multiple parameters must be estimated in the course of determining sample size. Incorporating sampling variability is discussed and illustrated in Chapter 3 and to a lesser extent in succeeding chapters.

If the observations are approximately normally distributed, then the range of possible observations divided by six will provide a reasonable estimate of σ. This assumes that the largest possible observation is at approximately $\mu + 3\sigma$ (the 99.865 percentile of a normal distribution) and the smallest value is at approximately $\mu - 3\sigma$ (the 0.135 percentile of a normal distribution). The difference between them is thus $(\mu + 3\sigma) - (\mu - 3\sigma) = 6\sigma$, so σ is estimated by Range/6. [Some writers have recommended that four be used as the divisor and others have indicated that either might be used (e.g., Lohr, 1999, p. 41), but six is the better choice, in my opinion, provided that there is sufficient information available regarding the largest possible value likely to be encountered as well as the smallest possible value. We will illustrate later how this can affect sample size determination.] Note that there is no sampling variability with the range method of estimating σ because the largest and smallest values likely to be encountered are assumed to be known.

Similarly, it is also possible to estimate σ if two percentiles of a normal distribution are known. For example, if the 90th percentile is believed to be 18 and the 20th percentile is considered to be 10, the 90th percentile is at $\mu + 1.28\sigma$ and the 20th percentile is at $\mu - 0.84\sigma$, with the difference between them thus being 2.12σ. So for this example we would estimate σ as $\hat{\sigma} = (18 - 10)/2.12 = 3.77$. [The software PASS can be used to estimate σ in this manner, as well as to estimate σ as (population range)/4.] Similarly, nQuery also has this capability. For example, if the option "Estimate standard deviation from sample percentiles" is selected in nQuery and "1" is entered for "Percentile" and 20 entered for the "lower observed value" and 80 entered for the "upper observed value," the software acts as if 1 is the population first percentile and 80 is the population 99th percentile, and further assumes that the population has a normal distribution. That is, it gives $\hat{\sigma} = 12.896$, which is obtained from $(80 - 20)/[2(2.32635)]$, with 2.32635 being the 99th percentile of the standard normal distribution and -2.32635 being the first percentile. Hand computation gives $\hat{\sigma} = 12.8957$, in agreement with the result given by nQuery.

Here process yield is being measured and let's assume that daily measurements are made, with past records indicating that almost all daily yield measurements have been between 91 and 109. Then we can estimate σ as $\hat{\sigma} = (109 - 91)/6 = 3.0$.

With σ estimated in this manner and approximate normality assumed, as stated, the test statistic is

$$Z = \frac{\bar{X} - 100}{3/\sqrt{n}}$$

Since we want $P(\bar{X} > 100 + 1.645\sigma/\sqrt{n})$, substituting $(100 + 1.645(3)/\sqrt{n})$ for $\bar{X}$ in the expression for Z of course produces 1.645, the threshold value for the test statistic, Z, such that any Z-value larger than 1.645 would result in rejection of the null hypothesis. Approximate normality for the distribution of $\bar{X}$ would have to be assumed in order to use Z, but in this example approximate normality of the individual observations was assumed in the estimation of σ. Approximate normality of the individual observations implies approximate normality of $\bar{X}$, and indeed the distribution of $\bar{X}$ will always be closer to a normal distribution than will the distribution of X (i.e., the individual observations).

Since 100 is the value for μ that forms the null hypothesis but 101 is assumed to be the true value of μ, we need to perform appropriate algebra starting from Eq. (2.1) to obtain an equation that will be used to solve for n. Subtracting 101 from both sides of the inequality and then dividing by $\sigma/\sqrt{n}$ produces

$$P\left(\frac{\bar{X} - 101}{3/\sqrt{n}} > 1.645 + \frac{100 - 101}{3/\sqrt{n}} \,\middle|\, \mu = 101\right) = .80$$

Thus,

$$1 - \Phi\left(1.645 + \frac{100 - 101}{3/\sqrt{n}}\right) = .80,$$

so that

$$\Phi\left(1.645 + \frac{100 - 101}{3/\sqrt{n}}\right) = .20$$

with $\Phi(\cdot)$ denoting the cumulative normal probability density function (*pdf*). Then

$$1.645 + \frac{100 - 101}{3/\sqrt{n}} = \Phi^{-1}(.20) = \Phi^{-1}(\beta) = -Z_\beta$$

with β denoting, in general, the probability of failing to reject the null hypothesis when it is false, with Z_β denoting the value of the standard normal variate with an area, β, to be specified, in the right tail of the distribution. (Later Z_α and $Z_{\alpha/2}$ will be used and the same general definition applies to these expressions.)

Thus, $\beta = 1 -$ power. Since $-Z_\beta = -Z_{.20} = -0.84$, the desired value of n is thus obtained by solving the equation

$$1.645 + \frac{100 - 101}{3/\sqrt{n}} = -0.84 \tag{2.2}$$

Doing so produces $n = [2.485*3/(101 - 100)]^2 = 55.58$. The value $n = 56$ would then be used so that the power is at least .80.

Of course, this result is conditional on $\sigma = 3$, so that .80 is the assumed power. What if the true value of the standard deviation was $\sigma = 2$? Then the actual power should be much greater than .80 since the smaller the standard deviation, the more powerful the test if everything else is kept constant. The power is

$$1 - \Phi\left(1.645 + \frac{100 - 101}{2/\sqrt{n}}\right) = .982$$

This large increase in the power should be expected since we are using a value for σ that is 33% smaller than the previous value.

What would the sample size have been if the range had been divided by 4 instead of by 6? Then $\hat{\sigma} = 4.5$ and following this same sequence of steps in PASS would lead to $n = 126$—more than twice the sample size when 6 was used as the divisor. How can there be such a huge difference? Since $n = 56$ was obtained from the calculation $n = [(2.485)(3)/(101 - 100)]^2$ and since $4.5 = 3(1.5)$ and the 1.5 would be squared in the formula, this means that the sample size obtained using $\hat{\sigma} = 4.5$ will be $(1.5)^2 = 2.25$ times the sample size obtained using $\hat{\sigma} = 3.0$. Here one could verify that 2.25(56) does equal 126. This shows how very sensitive the sample size is to the estimate of σ! Therefore, it is imperative that a "good" and perhaps conservative estimate of σ be used. This is pursued in Chapter 3, especially Section 3.3, in discussing how data from a pilot study might be used in determining sample size.

If we write Eq. (2.2) in the general form

$$Z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} = -Z_\beta \tag{2.3}$$

with μ denoting the actual mean and μ_0 denoting the hypothesized mean, we can see by performing some simple algebra that the general form of the solution for n for a one-sided test is

$$n = \left[\frac{(Z_\alpha + Z_\beta)\sigma}{(\mu - \mu_0)}\right]^2 \tag{2.4}$$

Notice from Eq. (2.4) that n is inversely related to the difference between the actual value of μ and the hypothesized value. This is quite intuitive because if the true value is much greater than the hypothesized value, that should not be hard to detect, so a large sample size would not be needed. Notice also that when the other components are fixed, the required sample size varies directly with the specified power, as 1.28 would be used in Eq. (2.4) instead of 0.84 if the specified power had been .90. Then $Z_\alpha + Z_\beta$ would have been 2.925 instead of 2.485. Of course, this is also intuitive because we would expect to need a larger sample size to have greater power in detecting a given difference between the hypothesized value and the true value.

Regarding the relationship between sample size and power, if we start from Eq. (2.3) and solve for power as a function of sample size, we obtain

$$\Phi\left[Z_\alpha - \frac{\sqrt{n}(|\mu - \mu_0|)}{\sigma}\right] = \beta$$

so

$$\text{Power} = 1 - \Phi\left[Z_\alpha - \frac{\sqrt{n}\,(|\mu - \mu_0|)}{\sigma}\right] \tag{2.5}$$

If the sample size is too small, then the fraction in the brackets is too small and the bracketed expression is too large, so that the power is then too low. To illustrate, let $\delta = |\mu - \mu_0| = 2$, $Z_\alpha = 1.645$, $\sigma = 8$, and $n = 90$. We will assume that the sample size was underestimated by (only) 10%, so that a sample size of 100 should have been used. The power, using either hand calculation or software such as PASS, can be shown to be .7663. If $n = 100$ had been used, then the power would have been .8038. If the user insists on having power of at least .80 in every application, then the underestimation of sample size for this example would be a problem.

It should be apparent from this example with slight sample size underestimation and from the results of Table 2.1 that the use of an internal pilot study of an appropriate size is very important. We will return to this topic with more examples in later chapters. ■

2.2.1 Bayesian Approaches

Bayesian sample size determination methods were reviewed in a general way at the beginning of this chapter. How would a Bayesian statistician or a person who wishes to use a Bayesian approach have proceeded relative to Example 2.1? Prior distributions are used in Bayesian statistics and here it is necessary to assign a prior probability to μ_0 and to a value for μ for which power is to specified. In the current example, μ_0 and μ are 100 and 101, respectively. A simple assignment of prior probabilities would be to use 1/2 for each of the two values. As discussed by Inoue et al. (2005), a decision regarding the null hypothesis under the Bayesian approach would be based on the posterior probabilities for each of the two values. (Posterior probabilities are computed by combining the prior probabilities with sample data. In sample size determination, data from a pilot study might constitute the Sample data.") A decision rule might be to not reject the hypothesized value if its posterior probability is at least $1/(1 + k)$ for a suitably chosen value of k. Let $\pi = P(\mu = \mu_0)$ and $\delta = \mu - \mu_0$. Bayesian sample size determination is then made using (π, K, δ, σ).

Inoue et al. (2005) gave an example with $\sigma = 1$, $\delta = .10$, $\alpha = .05$, and $\beta = .10$. The use of Eq. (2.4) produces $n = 856.385$, so $n = 857$ would be used. They showed how the same sample size is obtained with the Bayesian approach if the "minimum rate of correct classification" (i.e., the probability of making the right decision) is .9283. Of course, an experimenter would be more apt to select a number such as .80, .90, or .95 rather than .9283, so sample sizes obtained with frequentist and Bayesian approaches will generally differ.

What if σ could not be estimated in the preceding manner? The effectiveness of that method depends on (a) approximate normality of the individual observations and (b) information on the range of possible values. [What was tacitly assumed in obtaining the estimate of σ in Example 2.1 is that approximately 27 of every 10,000 observations is outside the interval (91, 109).] We would generally expect that information would be available to construct such an interval. If not, the process engineer and other workers involved in a manufacturing process or whatever type process was involved would presumably be able to provide a reasonable estimate of the process standard deviation. Therefore, a Bayesian approach, specifying some prior distribution for σ, should usually be unnecessary.

De Santis (2007) considered the use of data from similar studies in discussing a Bayesian approach to sample size determination but that could become a bit tricky because the populations and study conditions would need to be identical, or at least very similar.

Sahu and Smith (2006) considered a Bayesian approach to sample size determination, motivated by a practical application in clinical trials and in a financial audit. See also the Bayesian approach for sample size determination proposed by Berg (2006) for auditing property value appraisals to determine whether state accuracy guidelines are met.

Walker (2003) considered a Bayesian nonparametric approach that utilized decision theory and specifically solved for the sample size using the "maximization of expected utility." The author argues for a nonparametric approach because "a sample size is advocated without the benefit of observed data." The advocated approach is rather involved and complicated, however, and accordingly is likely to be primarily only of academic interest. (Sample size determination for nonparametric methods is covered in Chapter 10.)

Bayesian sample size determination methods for clinical trials were reviewed by Pezeshk (2003). This paper and many other papers on Bayesian sample size determination methods for clinical trials are discussed briefly in Section 7.8.

2.2.2 Probability Assessment Approach

As discussed by, for example, Chow et al. (2008, p. 18), for rare events, it may not be appropriate to determine sample size based on power, because the necessary sample size may be impractically large and small changes may not be of practical interest. Therefore, it may be more practical to determine sample size based on a probability statement that, say, a treatment group sample mean will be less than/greater than a control group sample mean. (Of course, no probability statement can be made about population means, at least with a frequentist approach, since population means are not random variables.) See Section 1.3.4 of Chow et al. (2008) for further details.

2.2.3 Reproducibility Probability Approach

This approach is due to Shao and Chow (2002) and relates to the fact that the U.S. Food and Drug Administration (FDA) usually requires at least two well-controlled clinical trials for providing evidence of the effectiveness and safety of a new drug. The general idea is to compute the sample size for the second trial using the concept of reproducibility probability. The sample size for the second trial is a function of the value of the test statistic for testing the equality of the treatment group mean (μ_2) and the control group mean (μ_1) that produces the desired reproducibility probability, as well as ϵ and C, with the difference in the population means assumed to change from $\mu_1 - \mu_2$ in the first trial to $\mu_1 - \mu_2 + \epsilon$ in the second trial, and the assumed common population variance assumed to change from σ^2 in the first trial to $C^2\sigma^2$ in the second trial, with $C > 0$. See Shao and Chow (2002) and/or Section 1.3.5 of Chow et al. (2008) for details.

2.2.4 Competing Probability Approach

Rahardja and Zhao (2009) proposed a new approach to sample size determination, which they termed the *competing probability* (CP) *approach,* with particular applications in clinical trials. Accordingly, the method is discussed and illustrated in Section 7.7. Briefly, the authors stated that “CP can be interpreted as the probability of the experimental treatment being more efficacious than the control treatment.” Thus, the focus is on this probability rather than a specified minimum effect size that an experimenter wishes to detect with a hypothesis test. See Lautoche and Porcher (2007) for somewhat related work.

2.2.5 Evidential Approach

Another method of determining sample size is to use an “evidential framework,” as described by Strug, Rohde, and Corey (2007). This utilizes the likelihood function, with evidence about a parameter classified as strong, weak, or misleading. See also Sutton, Cooper, Jones, Lambert, Thompson, and Abrams (2007).

2.3 FINITE POPULATIONS

In determining sample size it is generally sufficient to act as if the population is infinite and ignore the *finite population correction* (fpc) *factor,* which is *(N − n)/(N −* 1), with *N* denoting the population size. More specifically,

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n}\left(\frac{N-n}{N-1}\right) \tag{2.6}$$

Population sizes are unknown in most applications, so the fpc is generally ignored. In some applications, however, such as in accounting when the total number of records for a certain time period is known but it is too time consuming and costly to examine very record, the population size is known. If the sample size is to be a sufficiently large percentage of the population size (the usual rule of thumb is 5%), the fpc should be used in determining the sample size, whether a hypothesis test or a confidence interval is to be used.

When PASS is used for a hypothesis test of a single mean, the fpc is automatically used whenever the user specifies the size of the population rather than indicating that the population is infinite.

The use of the fpc in determining sample size for confidence intervals is discussed and illustrated in Section 2.4.1.

2.4 SAMPLE SIZES FOR CONFIDENCE INTERVALS

Although the relationship between confidence intervals and hypothesis tests was discussed briefly in Section 1.2, sample size could not be determined using a confidence interval so as to provide a specified power of the corresponding hypothesis test simply because power is not involved in confidence interval construction. As noted in Section 1.2, a confidence interval could be used to test a hypothesis and in some ways this is better than using a standard hypothesis test. (Sample size could be computed using a hypothesis test approach and then a confidence interval or one-sided confidence bound constructed rather than performing the hypothesis test, but that would be a mixture of a hypothesis testing approach and a confidence interval approach.) While stating a trend away from hypotheses tests and toward confidence intervals in epidemiologic analyses, Greenland (1987) notes that confidence intervals can be misleading indicators of the discriminatory power of a study if they are not properly centered and Jiroutek, Muller, Kupper, and Stewart (2003) presented what they claimed to be improved methods of determining sample size for confidence intervals. Specifically, with the term "Validity" defined as a confidence interval that contains the true parameter value, they addressed the question "Given validity, how many subjects are needed to have a high probability of producing a confidence interval that correctly does not contain the null value when the null hypothesis is false and has a width no greater than δ?"

Sample size can be determined so as to give a confidence interval of a specified width, or equivalently, a maximum error of estimation, which is just the halfwidth of the confidence interval. Bristol (1989) discussed the determination of sample size for confidence intervals of a specified width and compared that to the usual determination of sample size for hypothesis tests with a stated power.

To illustrate, consider a confidence interval for μ. The expression for the width of a confidence interval for μ using Z is obtained by taking the difference between the upper limit (*U.L.*) and the lower limit (*L.L.*). Specifically, the width, *W*, is given by

$$\begin{aligned} W &= L.L. - U.L. = \bar{X} + Z_{\alpha/2}\sigma/\sqrt{n} - (\bar{X} - Z_{\alpha/2}\sigma/\sqrt{n}) \\ &= 2Z_{\alpha/2}\sigma/\sqrt{n} \end{aligned}$$

Thus, $W = 2Z_{\alpha/2}\sigma/\sqrt{n}$. Solving this equation for n gives $n = (2Z_{\alpha/2}\sigma/W)^2$. Thus, *n* is inversely related to *W*, with a small value of *W*, relative to σ, producing a large value of *n*. Of course, an estimate of σ is needed, perhaps to be obtained using the method in Example 2.1, or from prior data such as historical records.

Since $\hat{\mu} = \bar{X}$ and the latter is in the middle of the confidence interval, this means that the maximum error of estimation is $E = W/2$, with probability $1 - \alpha$. Thus, sample size could be computed for a specified maximum error of estimation, with the expression for *n* being

$$n = (Z_{\alpha/2}\sigma/E)^2 \tag{2.7}$$

with *n* rounded to the next integer value since the result will almost certainly not be an integer. The choice between the use of *E* or *W* of course depends on the focus of the experimenter. We can see from Eq. (2.7) that if an experimenter has a value of *E* in mind but decides to switch to a value half as large, then the necessary sample will, on average, be four times the required value with the original value of *E*. Thus, extra precision could be quite expensive and there is no guarantee of what will happen on individual samples since $\bar{X}$ is a random variable. Thus, there is randomness involved even if σ were assumed to be known. This motivated Webb, Smith, and Firag (2010) to look at the probability of achieving improved accuracy with an increased sample size, extending the results of Gauch (2006). The results of Webb et al. (2010) showed that very large increases in the sample size are required in order for the probability that the error of estimation with the increased sample size, which they term the "gain probability," will be close to 1. (The gain probability is the probability that the estimator based on the larger of two sample sizes is closer to the true parameter than the estimator based on the smaller sample size.) Webb et al. (2010) stated that "it is known that setting sample size requirements to guarantee improved accuracy with a given probability can lead to an alarmingly high demand on sample size" and they quantified that statement, although they did so using graphs rather than tables. Their Figure 8 combined with some statistical work shows that the gain probability is well-approximated by $0.627 + 0.0751[\log(k/n)]$, assuming normality and essentially independent of *n*, with *k* denoting the number of additional samples and $2 \leq k/n \leq 10$. Thus, when $k/n = 10$ so that the sample size is being increased by 10 times the original sample size, the gain probability would be approximated by .80, whereas their Figure 8 suggests that the probability is approximately .81. It should be of concern that such an enormous increase in the sample size results in a gain probability that is considerably less than 1.

When viewed from a cost perspective, which Webb et al. (2010) did not consider, in addition to statistical considerations, this would likely be a very undesirable way of determining sample size.

The user does not have a choice when PASS is used for determining the sample size, however, as the user specifies the "distance from mean to limit(s)," which of course is E. To illustrate, let $\sigma = 2, E = 1$, an infinite population is assumed, and a 95% two-sided confidence interval for the mean is desired. The software gives $n = 16$ as the required sample size. We can see that hand computation would give $n = [1.96(2)]^2 = (3.92)^2 = 15.37$, so $n = 16$ would be used, as the software indicated.

Kupper and Hafner (1989) reported that this formula performed poorly in their numerical work. This requires some clarification. The sample size expression in Eq. (2.7) is an exact result *if* the individual observations have a normal distribution and if the population standard deviation, σ, is *known.* Of course, neither condition is likely to be met in practice. Kupper and Hafner (1989) stated that the use of $n = \left(Z_{\alpha/2}\sigma/E\right)^2$ "always leads to a serious underestimation of the required sample size. This surprising result can be illustrated by appealing to an overlooked, but nevertheless important, result due to Guenther (1965)." The result is really not surprising because if σ is later assumed unknown after solving for n for known σ, the sample size so determined will be too small since the variability resulting from using s to estimate σ will not have been accounted for! The reader is asked to show in Exercise 2.1 that this sample size formula does work as intended, when the *conditions for its use are met.* The discussion of Kupper and Hafner (1989) is not irrelevant, however, as there is at least one software package that solves for n using an inputted value of σ in the two-sample case, then allows the user to specify whether the test statistic to be used is for a pooled t-test (i.e, unknown common σ) or not. This is discussed further in Section 3.2. There is a definite inconsistency when software is used and any type of t-test is selected, for one sample or more than one sample, because σ must be indicated as known to solve for the sample size, but if we knew σ we wouldn't be using t in the first place! An important contribution of Kupper and Hafner (1989) is that they emphasized the need to use sample size formulas for simple confidence intervals that incorporate the appropriate expression for factoring in a tolerance or coverage probability for the estimation of σ by S. This is important because sample size determination software such as PASS assume that the value for the standard deviation that the user enters will be that value in all future samples when the menu item with the coverage probability is not selected. Obviously all future samples will not have the same standard deviation.

Note that this assumed two-sided confidence interval does not correspond directly to Example 2.1 because that was a one-sided hypothesis test, which corresponds to a one-sided confidence bound, not to a two-sided confidence interval. It is not possible to solve for n for a one-sided confidence bound of a desired width because such bounds of course do not have a finite width since no value is used for a bound on the other side.

We could, however, do something similar and solve for n such that there is a maximum error of estimation in the direction of interest. For a two-sided confidence interval, the maximum error of estimation (with the stated probability) is half the width of the confidence interval. For a one-sided confidence bound, the maximum error of estimation is given by the expression for the estimator and the expression for the confidence bound.

For example, assume normality for the sake of illustration, known σ, and that μ is to be estimated by $\bar{X}$ such that $\bar{X}$ should not exceed μ by more than two units with probability .95. Stated differently, we want the lower confidence bound on μ to be within two units of $\bar{X}$ with probability .95. Since the 95% lower confidence bound on μ is given by the expression $\bar{X} - 1.645\sigma/\sqrt{n}$, $\bar{X} - \left(\bar{X} - 1.645\sigma/\sqrt{n}\right) = 1.645\sigma/\sqrt{n}$. Since this difference is to be at most 2, $1.645\sigma/\sqrt{n} \le 2$. Solving this inequality gives $n \ge 0.676\sigma^2$. Thus, if $\sigma = 10$, for example, $n = 68$ would be used.

Of course, normality does not exist in practice, so the user who wants to solve for the sample size to give a confidence interval of a specified width or a maximum error of estimation with a certain probability must also contend with nonnormality and consider how large the sample size would have to be to overcome the fact that normal theory methods are being used on nonnormal data. Boos and Hughes-Oliver (2000) addressed this issue and gave some recommendations that were based on the population skewness. They concluded, however, by stating that “These are rough generalizations and we encourage readers to find their own rules.”

2.4.1 Using the Finite Population Correction Factor

As indicated in Section 2.2, it is sometimes necessary to use the finite population correction factor (fpc), which is $(N - n)/(N - 1)$, although some sources use $(N - n)/(N)$. Whichever quantity is used, it is multiplied times $\sigma_{\bar{x}}^2$ when $\bar{X}$ is the random variable that is used, as in estimating a population mean or total.

Of course, in solving for the sample size we won’t know whether or not the fpc will be needed unless we know what fraction of the population will have to be sampled in order to obtain a desired confidence interval width or to reject the null hypothesis when the parameter being tested has a certain value.

To illustrate, let’s assume that σ has been estimated by the range method, with the estimate being 2.0, so Z will be used in constructing a 95% confidence interval that is to have width $W = 1$, and $N = 1000$. From the development in Section 2.3 combined with Eq. (2.5), we have $2(Z_{\alpha/2})\sigma_{\bar{x}} = W$, so that

$$2(Z_{\alpha/2})\sqrt{\frac{\sigma_x^2}{n}\left(\frac{N-n}{N-1}\right)} = W \tag{2.8}$$

Solving for n produces

$$n = \frac{4NZ^2_{\alpha/2}\sigma^2}{(N-1)W^2 + 4Z^2_{\alpha/2}\sigma^2} \tag{2.9}$$

Using $\hat{\sigma} = 2.0$, W = 1, $N = 1000$, and $Z_{\alpha/2} = 1.96$, we obtain $n = 57.96$, so $n = 58$ would be used. If the fpc had not been used, the solution would have been $n = 4Z^2_{\alpha/2}\sigma^2/W^2 = 61.47$, so $n = 62$ would have been used. Thus, there is a slight difference in this case. Here *n/N* was only .058, so it isn't much larger than .05. There would have been a greater difference in the two sample sizes if *n/N* had been larger.

If we wanted the sample size expression in terms of the maximum possible error of estimation, *E,* we would use the relationship between *W* and *E* noted in Section 2.4 (*E* = *W*/2), and substitute 2*E* for *W* in Eq. (2.9). Doing so produces

$$n = \frac{NZ^2_{\alpha/2}\sigma^2}{(N-1)E^2 + Z_{\alpha/2}\sigma^2} \tag{2.10}$$

Assume that a company wishes to estimate its average accounts receivable at some point in time. The company's computerized accounting system is not set up to easily generate this number, and since there are too many accounts to look at in order to do this manually, a sample will be taken. It might be necessary to stratify the population in applications such as this when there is a wide range of amounts, as a smaller sample size will be possible by sampling within well-determined strata rather than sampling at random from the population.

There are skeptics who believe that sample size determination in auditing is not a worthwhile endeavor. In particular, Wilburn (1984, p. 47) stated: "Any meticulous, absolutely exact or time consuming procedure with precisely determined sample size is neither justifiable nor desirable in most audits. Predetermined sample sizes are generally based on assumptions which may not be applicable in the audit circumstances." Of course, the last sentence applies to virtually *any* sampling situation, not just to auditing. There is evidence from documents available on the Internet, however, that statistical methods of determining sample size have fallen into disfavor with the auditing profession. Even though there are problems inherent in the statistical approach to sample size determination, worse results could easily result when a nonstatistical approach is used.

Nevertheless, there appears to be the growing realization that sample size determination is fraught with problems. One sign of this is the recent discussion in Noordzij, Dekker, Zoccali, and Jager (2011). In particular, they stated:

> These examples show the most important drawback of sample size calculations; investigators can easily influence the result of their sample size calculations by changing the components in such a way that they need fewer patients, as that is usually what is most convenient to the researchers. For this reason, sample size calculations are sometimes of limited value. Furthermore, more and more experts are expressing criticism of the current

> methods used. They suggest introducing new ways to determine sample sizes, for example, estimating the sample size based on the likely width of the confidence interval for a set of outcomes.

Bland (2009) has argued in favor of determining sample size based on the width of a confidence interval, not the power of a hypothesis test. Such an approach would have considerable merit, but this probably won't happen to any great extent in practice until software for sample size determination moves in that direction.

2.4.1.1 Estimating Population Totals

In addition to means and other parameters, sampling books also cover the estimation of population totals. In order to estimate a population total, a population must be finite and known. We have used *E* to denote the maximum error of estimation in estimating μ by $\bar{X}$. Since a population total would be estimated by $N\bar{X}$, with *N* denoting the population size, it would be logical to use *NE* to represent the maximum error in estimating the corresponding population total.

Using *NE,* the sample size would be the same as the sample size for estimating μ with error *E.* This should be intuitively apparent, but we can also easily show it, as follows, with need for the fpc assumed. Since

$$\text{Var}(N\bar{X}) = N^2\text{Var}(\bar{X}) = N^2 = N^2\frac{\sigma_x^2}{n}\left(\frac{N-n}{N-1}\right)$$

then

$$\sigma_{N\overline{X}} = N\sqrt{\frac{\sigma_x^2}{n}\left(\frac{N-n}{N-1}\right)}$$

The sample size determination as a function of *E* would then be obtained by solving the equation

$$Z_{\alpha/2}N\sqrt{\frac{\sigma_x^2}{n}\left(\frac{N-n}{N-1}\right)} = NE \tag{2.11}$$

Since $E = W/2$, it can easily be seen that Eq. (2.11) is equivalent to Eq. (2.8).

Of course, a sample size formula could also be derived independent of the error expression for the population mean. Let *E** denote the selected maximum error of estimation for the population total. The sample size expression would then be obtained by substituting *E** for *NE* in Eq. (2.11) and then solving for *n.* Doing so produces

$$n = \frac{NZ_{\alpha/2}^2\sigma^2}{(N-1)(E^*)^2 + Z_{\alpha/2}^2\sigma^2}$$

It can be noted that it is never wrong to use the fpc in deriving the sample size expression for a finite population, as was done in deriving this expression. It is wise to use it whenever the population size is known because the effect that it has can't be determined until the value of n is obtained. If n/N is quite small, the use of the fpc in deriving the formula will have little effect; otherwise, the effect will be noticeable.

2.5 CONFIDENCE INTERVALS ON SAMPLE SIZE AND POWER

Since sample size is computed using an estimate of at least one other parameter, such as a variance being estimated, this means that the resultant sample size is a random variable, as is the assumed power. Therefore, a more realistic approach would be to construct confidence intervals for sample size and power, rather than assume that these are known. Although this is not built into sample size determination software and thus may be seldom used, the interested reader is referred to Taylor and Muller (1995).

2.6 SPECIFICATION OF POWER

There is no reason why .80 should be used, in general, for sample size calculations. Indeed, in statistics we generally prefer more "certainty" than that, such as a 95% or 99% confidence interval. Cohen (1988, p. 55) addressed the selection of a value for power and stated that "in the judgment of the author, for most behavioral science research (although admitting of many exceptions), power values as large as .90–.99 would demand sample sizes so large as to exceed an investigator's resources." Power depends on a number of factors, however, including the size of effect to be detected with high probability, and the standard deviation of the random variable that serves as the estimator of the parameter being tested. Thus, power values of at least .90 won't necessarily result in an impractically large sample size.

In pointing out problems with the assumptions and mechanics of sample size computations, Parker and Berman (2003) made a very valid point when they stated the following: "First, there is no a priori reason why one specific value of a difference . . . is worthy of detection with a certain power, while a slightly different value is worthy of less (or more) power of being detected." They feel that emphasis should shift from determining the sample size that is needed to detect a particular difference with a specified power to determining the information that is gained from using a particular sample size. Of course, we would ideally like to be able to detect the true value of the parameter with a high probability, rather than have a high probability of detecting a parameter value that might be somewhat arbitrarily chosen.

Their points are well taken. If we test that a population mean is 50 but a true value of μ between 49.5 and 50.5 is "close enough," why should there be a higher probability of rejecting the null hypothesis when $\mu = 50.4$ compared to $\mu = 50.3$ when both differences from 50 are deemed inconsequential, and thus the difference between 50.4 and 50.3 is also inconsequential? Debates such as this argue indirectly against the use of hypothesis testing, in general, and thus indirectly argue against the determination of sample size in a hypothesis testing context.

2.7 COST OF SAMPLING

Regardless of what methods are used to determine sample size, the cost of sampling must be considered. Remember that although increasing the sample size beyond what was originally envisioned will give increased power, it will also generally result in increased costs. So the marginal cost must be considered relative to the gain in precision/power. See `http://www.pmean.com/08/TooMuchPower.html` for a description of a study in which costs were considered, and note the following statement of Simon (2008): "The optimal sample size is one where the incremental value of improved precision is offset by the direct and indirect costs of obtaining an additional patient. No one does it this way, but they should." Of course, this assumes that the optimal sample size results in a cost that is within the budget for a study. If not, there is not much point in identifying a point of diminishing returns. See Chapter 7 of Brush (1988) for information on using costs and loss functions in the determination of sample size.

In addition to the explicit cost of sampling, it is recognized that sampling can simply be burdensome, especially when large samples are obtained. This point was made, for example, in the *2006–2007 Quality Assurance Program Sampling Guide* for Federal Student Aid, as educational institutions were required to sample 350 student records only every other year, rather than every year.

2.8 ETHICAL CONSIDERATIONS

Referring to clinical studies, various authors (e.g., Halpern, Karlawish, and Berlin, 2002) have claimed that many such studies do not have a power of at least .80 for detecting a minimum important effect, as discussed by Bacchetti, Wolf, Segal, and McCulloch (2005). [Readers may be interested in the response to that article by Halpern, Karlawish, and Berlin (2005).] See also Maxwell (2004) for a discussion of low power in psychological studies.

The argument has been that it is unfair to ask study participants to accept the risks and discomfort of being participants if the study does not have sufficient power to detect an effect of the minimum size that is considered to be important. Indeed, in referring to clinical trials, Altman (1980) stated: "If the sample size is too small there is an increased risk of a false-negative reading." See also the discussion of this in Lenth (2001).

That argument has received some criticism, however, as others have stated that such studies may still produce useful point estimates and confidence intervals, or contribute to meta-analyses. The latter was the point made by Chalmers, Levin, Sacks, Reitman, Berrier, and Nagalingham (1987). (Meta-analyses of clinical trials does present certain challengers, however, as discussed in Section 7.6.) Interestingly, Schulz and Grimes (2005) stated that Tom Chalmers considered Freiman, Chalmers, Smith, and Kuebler (1978) to be the most damaging paper he co-authored. Why? The paper was heavily cited (over 600 citations) and took the position that trials with low power were unethical. Many people were influenced by that paper and adopted the same position.

Bacchetti et al. (2005) have a different counterargument, explaining that "the balance between a study's value and the burdens accepted by its participants does not improve as the sample size increases. Thus, the argument for ethical condemnation of small studies fails even on its own terms." The authors conclude: "Indeed, a more legitimate ethical issue regarding sample size is whether it is too large."

Study participants are divided into two (or more) groups, so if the "treatment group" is receiving something that is beneficial, then by definition the other group does not receive it. So the larger the sample size, the greater the number of people who are not receiving a treatment that they may need, or at least an alternative to that treatment. So ethical considerations do have to be made, at least in a nonquantitative manner, for certain types of studies when sample size determination is being made.

Sample size computations depend on whether the test is one-sided or two-sided. This can be seen from Eq. (2.4). If Example 2.1 had been a two-sided test instead of a one-sided test, Z_α in Eq. (2.4) would be replaced by $Z_{\alpha/2}$. Since $\alpha/2$ is obviously less than α, $Z_{\alpha/2}$ is larger than Z_α since the former is further out into the right tail of the standard normal distribution than is the latter. Assume, as in Example 2.1, that power is to be .80. For $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$ compared to $Z_\alpha = 1.645$. Since $(0.84 + 1.96)^2 = 7.84$, $(0.84 + 1.645)^2 = 6.175$ and 7.84/6.175 = 1.27, the computed sample size for a two-sided test would thus be 27% greater than the sample size for a one-sided test. Stated differently, the necessary sample size for a one-sided test would be 79% of the sample size needed for a two-sized test.

Despite the popularity of one-sided testing in clinical research, Moyé and Tita (2002) argued in favor of two-sided tests and gave a graph that plotted the percentage against α, with 79% being one of the points on the graph. One point they made to support their contention is that a treatment may be harmful rather than helpful, and that harm often does occur. Of course, it would be important to detect that, and to detect it as soon as possible. An important quote from their paper is: "Here, the finding of harm in a one-tailed test designed to find benefit makes it ethically unacceptable but scientifically necessary to reproduce the result. This conundrum causes confusion in the medical community and could have been completely avoided by carrying out a two-tailed test from the beginning." They also stated that "a one-tailed test designed exclusively to find benefit does not permit the assessment of the role of sampling error in producing harm, a dangerous omission for a profession whose fundamental tenet is to first do no harm."

Knottnerus and Bouter (2001) offered a counterargument, stating that more people would receive the inferior treatment (in, say, a clinical study), but a study might be terminated early if there is strong evidence of benefit, as pointed out by Moyé and Tita (2002).

These are some important points and other important points regarding the choice of one-sided versus two-sided tests were also made by Bland and Altman (1994). While indicating that a one-sided test is sometimes appropriate, they strongly favor two-sided tests in clinical research "unless there is a very good reason for doing otherwise." Their statement—"If a new treatment kills a lot of patients we should not simply abandon it; we should ask why this happened"—is clearly a sizable understatement! Their paper prompted Letters to the Editor by R. Wolterbeek and M. W. Enkin in the October issue of that journal, with each expressing a dissenting opinion. This was followed by a response from Bland and Altman, who made the important point that in clinical research there is a definite need to determine why the results were in the opposite direction from what would have been specified in the alternative hypothesis if a one-sided test had been used.

Of course, algorithmic approaches to sample size determination in software and applets do not incorporate ethical considerations, so the reader may want to study the paper by Bacchetti et al. (2005) for ways in which ethical considerations should influence sample size. Similarly, sampling costs are also generally not incorporated in software, so it is up to practitioners to assume the initiative. As Simon (2008) stated: "No one does it this way, but they should." Of course, it would be helpful if software would allow the input of costs, but I am not aware of any statistical software that has ever had this capability, although surely there must be some software, perhaps little known, that allows the user to input costs for determining sample size.

2.9 STANDARDIZATION AND SPECIFICATION OF EFFECT SIZES

Effects are generally given as the difference between the hypothesized value of a parameter and a value that an experimenter wishes to detect with a high probability, standardized in some manner. For example, the difference between the hypothesized value of μ and a detectable value may be standardized by dividing that difference by σ. Let that standardized statistic be represented by d. For the comparison of two means, Cohen (1988) defined "small," "medium," and "large" values of d as .10, .30, and .50, respectively. We will see in later chapters that such designation will generally be inappropriate, and the use of these labels has been criticized by Lenth (2001) and others. In particular, note that Lenth (2006–2009) refers to these as "T-shirt effect sizes" and lists this as one of "two very wrong things that people try to do with my software." While admitting that the determination of effect size in ecological studies is difficult due to the paucity of relevant data, Fox, Ben-Haim, Hayes, McCarthy, Wintle, and Dunstan (2007) agree with Lenth (2006–2009) that "T-shirt effect sizes . . . is not the way to resolve this problem." They proposed the use of "info-gap theory," which is intended "to address the 'robustness' of decision making under uncertainty."

Using standardized effects enables a practitioner to avoid thinking about the magnitude of σ, which is not a good idea. Regarding the "shirt sizes approach," Thomas (1997) stated: "These conventions are widely used in psychology and other disciplines, where a medium standardized effect size may correspond with the median effect size found in psychological research (Sedlmeier and Gigerenzer 1989)."

Regarding effect sizes, van Belle (2008, Chap. 2) stated: "Some social science journals insist that all hypotheses be formulated in terms of effect size. This is an unwarranted demand that places research in an unnecessary straight jacket."

Cohen (1988) gave a different set of numbers for other types of hypothesis tests; as indicated in his Section 10.2.2. In general, however, it is unwise to try to associate various degrees of change with specific numbers. How the magnitude of a change would be assessed should certainly depend on what is being measured as well as the field of application.

2.10 EQUIVALENCE TESTS

In this section we introduce sample size determination for equivalence tests and also consider sample size determination for noninferiority and superiority tests, following the introduction and comparison of these different types of tests in Section 1.5.

Equivalence testing was proposed when the objective is to show the "equivalence," appropriately defined, between two population means or proportions. Garrett (1997) stated that "equivalence tests are perhaps the second most useful general class of hypothesis tests after the standard hypothesis testing framework."

The essential difference between traditional hypothesis testing (THT) when two populations are involved and equivalence testing is that in THT the focus is on the null hypothesis, which is often equality ("equivalence"), and seeing if it can be rejected in favor of one method, treatment, and so on being better than the other one, whereas in equivalence testing the objective is to show equivalence, appropriately defined. (Recall that, except in testing distributional assumptions, the null hypothesis is what we doubt to be true.)

In equivalence testing, the form of the equivalence could be stated in either the null hypothesis or the alternative hypothesis, but regardless of how it is set up, we can never "prove" equivalence (nor prove any hypothesis, in general, with sample data), contrary to such wording that is used in the literature, which is really a misnomer. What *can* be done, however, is to determine whether or not there is "essential equivalence" relative to a range of scientific or clinical indifference.

That is, if we take a sample from each of two populations, we don't know if the parameters of interest are either equal or differ by a very tiny amount without the sample being the entire population, so that there would then be no sampling variability.

As discussed by Dixon (1998), much of the statistical development of equivalence testing has been motivated by the regulatory requirement that a newly developed generic drug must be shown to be equivalent to the corresponding name-brand drug. This leads to the consideration of bioequivalence and sample size determination for it (see, e.g., Phillips, 1990). With AUC denoting the plasma concentration time curve and $C_{\max}$ denoting the peak concentration, the FDA has considered a test product to be bioequivalent to a standard (reference) product if a 90% confidence interval for the geometric mean ratio of AUC and $C_{\max}$ between the test and reference products falls within the interval 80% to 125%. Alternatively, a hypothesis test could be performed using an upper bioequivalence limit of 1.25 and a lower bioequivalence limit of .80.

Equivalence testing has applications beyond drug testing, however, and Dixon (1998) indicated that the general principles can also be used in environmental applications. See also Blackwelder (1998), Diletti, Hauschke, and Steinijans (1992), and Chow and Liu (1999) for biostatistical applications, and see Friede and Kieser (2003) for considerations that should be made when an internal pilot study is used in conjunction with equivalence and noninferiority testing. O'Quigley and Baudoin (1988) is a paper on general approaches to bioequivalence and Zhang (2003) gave a simple formula for sample size calculation in bioequivalence studies. See also Ganju, Izu, and Anemona (2008). Readers interested in the derivation of sample size formulas for equivalence testing as well as for tests of equality, noninferiority, and superiority are referred, in particular, to Chow, Shao, and Wang (2002, 2008). See Pocock (2003) for a discussion of the pros and cons of noninferiority trials.

■ EXAMPLE 2.2

To illustrate the computation of sample size for equivalence testing, consider Example 2.1 but now let's assume that a process yield of 105 from a new process that requires less manual attention is "equivalent" to a process yield of (at most) 100. Perhaps the (daily) process yield in excess of 102 could not be easily utilized. That is, there is more or less a target for process yield. The null hypothesis might be set up so that the difference is greater than 2, with the alternative hypothesis being that the difference is at most 2. Thus, we would want to reject the null hypothesis.

We will assume, as in Example 2.1, that 3.77 is our estimate of σ and that the power is to be .80, with $\alpha = .05$. The necessary sample size would be computed as follows (see Chow, Shao, and Wang, 2008, p. 57), using a normal approximation approach.

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2 \sigma^2}{\delta^2}$$

with δ denoting the equivalence limit, which here is 2. Thus,

$$n = \frac{(1.645 + 0.84)^2 (3.77)^2}{(2)^2} = 21.94$$

so that $n = 22$ would be used. Note that this is a one-sample test whereas it was stated earlier in this section that equivalence testing was developed for testing the equivalence of two means or proportions. It can also be used for one-sample problems, as illustrated in this SAS Software documentation (`http://support.sas.com/documentation/cdl/en/anlystug/58352/HTML/default/viewer.htm#chap12_sect4.htm`). ■

See also a simple method for sample size determination that was given by Zhang (2003). For additional information on equivalence testing, see Wellek (2010).

2.11 SOFTWARE AND APPLETS

Although sample size determination and power computation formulas are available for hand computation, if desired, it is preferable to use either commercial software or Java applets (i.e., freeware), or a combination of the two. Therefore, in subsequent chapters the formulas are given, where appropriate, but there is also considerable discussion of available software (and their strengths), as well as applets. Sample size determination software and applets vary greatly in terms of overall capabilities. For example, some software offer an option for a continuity correction when a normal approximation is applied to a discrete random variable, whereas other software do not.

Although there is considerable software discussion (and some illustration) in the succeeding chapters, the discussion is limited to a relatively small number of software and a few applets. A broader discussion of software is given by Dattalo (2008), which includes recommendations of which software to use for each type of test that is presented. It is obvious that Dattalo (2008) has a preference for freeware, which is evidenced by the statement (p. 13): “Readers should be aware that, whenever possible, the approach recommended here is to estimate sample size with GPower, which is a free power analysis program."

Certainly freeware and Java applets do have a role to play, especially in education. This is discussed in some detail by Anderson-Cook and Dorai-Raj (2003), who have a link to a nice applet (http://www.amstat.org/publications/jse/v11n3/java/Power/Power3Applet.html) that can be used to see with dynamic graphics the relationship between power and the value specified for the alternative hypothesis for one-sample and two-sample tests of means and proportions. Lenth's applet (http://www.cs.uiowa.edu/∼rlenth/Power) is the most popular and best known general purpose applet for sample size determination and power. The Web page http://statpages.org/#Power also has considerable capabilities and note the long list of Web pages at http://www.webstatschecker.com/stats/keyword/sample_size_calculation_power. Applets are great for students and people engaged in self-study who may not have access to software for sample size determination.

Regarding the software PASS that was mentioned in Section 2.1, Dattalo (2008, p. 13) stated: "For researchers who prefer a comprehensive statistical package, PASS is recommended. PASS is capable of providing a wide range of sample size calculations."

In addition to PASS, Power and Precision, nQuery, and SiZ will be discussed and illustrated in subsequent chapters, and there will also be references to other software, including Stata and SAS Software. Users of the latter who want to employ it for sample size determination may want to start with the overview at http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#intropss_toc.htm. The sample size determination capabilities of SAS Software are somewhat limited compared to the capabilities of certain software that is solely for sample size determination and power, so users will need to write short SAS Software programs to match those capabilities. There is some discussion of this in subsequent chapters. See http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_clientpss_sect002.htm for a list of the sample size and power capabilities of Version 9.22 using the SAS Power and Sample Size (PSS) application, and also see Watson (2008). See also Zhao and Li (2011), who illustrated how to determine sample size using simulations in SAS Software for models that are more complicated than the models that are handled by the POWER procedure in SAS Software.

Stata also has sample size determination, primarily with its sampsi command and also with the stpower command, which computes power and sample size for survival analysis (see Chapter 9). Unlike MINITAB, however, Stata has limited sample size determination capability and does not have a menu mode option, whereas MINITAB does have such an option.

Regarding freeware, probably the best-known and most often used freeware for sample size computations is Lenth's applet (http://www.cs.uiowa.edu/~rlenth/Power/), although Dattalo (2008, p. 13) expresses a preference for G*Power (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3). See also Faul, Erdfelder, Lang, and Buchner (2007). Users of R may be interested in the code for sample size determination given in Cohen and Cohen (2008).

Although G*Power does have considerable capabilities, especially for freeware, PASS 11 can do much more. In particular, PASS has capability for sample size determination in quality control; such capability is not available in any other software. This is discussed further in Chapter 8.

2.12 SUMMARY

Determining sample size will generally be difficult because of the necessity of having to specify values for unknown parameters, such as σ, or having to specify the change that an experimenter wishes to detect in standard deviation units when it is more natural for an experimenter to think in terms of the unit of measurement. As when any statistical tool is used, the results will be only as good as the assumptions that are made.

These types of problems do not render sample size determination useless, but experimenters should keep in mind that the specified power used in sample size determination is not going to be the actual power that a study has. (There is always uncertainty in statistics because random variables are involved.)

REFERENCES

Adcock, C. J. (1988). A Bayesian approach to calculating sample sizes. *The Statistician*, **37**, 433–439.

Adcock, C. J. (1997). Sample size determination: A review. *The Statistician*, **46**(2), 261–283.

Altman, D. (1980). Statistics and ethics in medical research, III: How large a sample. *British Medical Journal*, **281**, 1336–1338.

Anderson, K. M. (2006). Adaptive designs. Sample size-reestmation: A review and recommendations. Slide presentation in the FDA/Industry Statistics Workshop, PhRMA Adaptive Designs Working Group, Philadelphia, October 27. (www.amstatphilly.org/events/fall/2006/KeavenSSR.ppt)

Anderson-Cook, C. M. and S. Dorai-Raj (2003). Making the concepts of power and sample size relevant and accessible to students in introductory statistics using applets. *Journal of Statistics Education*, **11** (electronic journal).

Ariola, M. M. (2006). *Principles and Methods of Research*. Manila, Phillipines: Rex Book Store, Inc.

Bacchetti, P. (2010). Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine,* **8**, 17. Rejoinder: Good intentions versus CONSORT's actual effect.

Bacchetti, P., S. G. Deeks, and J. M. McCune (2011). Breaking free of sample size dogma to perform innovative translational research. *Science Translational Medicine*, **3**(87), 24.

Bacchetti, P., C. E. McCulloch, and M. R. Segal (2008). Simple, defensible sample sizes based on cost efficiency. *Biometrics*, **64**, 577–585.

Bacchetti, P., L. E. Wolf, M. R. Segal, and C. E. McCulloch (2005). Ethics and sample size. *American Journal of Epidemiology*, **161**(2), 105–110. Discussion: pp. 111–113.

Berg, N. (2006). A simple Bayesian procedure for sample size determination in an audit of property value appraisals. *Real Estate Economics*, **34**, 133–155.

Betensky, R. A. and C. Tierney (1997). An examination of methods for sample size recalculation during an experiment. *Statistics in Medicine*, **16**(22), 2587–2598.

Binkett, M. A. and S. J. Day (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine*, **13**, 2455–2463.

Blackwelder, W. C. (1998). Equivalence Trials. In *Encyclopedia of Biostatistics*, Vol. 2. New York: Wiley, pp. 1367–1372.

Bland, J. M. (2009). The tyranny of power: Is there a better way to calculate sample size? *British Medical Journal*, **339**, 1133–1135.

Bland, J. M. and D. G. Altman (1994). One- and two-sided tests of significance. *British Medical Journal*, **309**, 248. Response as Letters to the Editor by R. Wolterbeen and M. W. Enkin, with response by Bland and Altman, **309**, 873–874. (This is free to registered users; see `http://www.bmj.eom/cgi/content/full/309/6958/873/a`.)

Boos, D. D. and J. M. Hughes-Oliver (2000). How large does *n* have to be for Z and *t* intervals? *The American Statistician*, **54**(2), 121–128.

Bristol, D. R. (1989). Sample sizes for constructing confidence intervals and testing hypotheses. *Statistics in Medicine*, **8**, 803–811.

Browne, R. H. (1995). On the use of a pilot sample for sample size determination.*Statistics in Medicine*, **14**, 1933–1940.

Brush, G. G. (1988). *How to Choose the Proper Sample Size*. Volume 12: The ASQC Basic References in Quality Control: Statistical Techniques (J. A. Cornell and S. S. Shapiro, eds.). Milwaukee, WI: American Society for Quality Control.

Chalmers, T. C., H. Levin, H. S. Sacks, D. Reitman, J. Berrier, and R. Nagalingham (1987). Meta-analysis of clinical trials as a scientific discipline, I: Control of bias and comparison with large co-operative trials. *Statistics in Medicine*, **6**, 315–328.

Chow, S. C. and J. P. Liu (1999). *Design and Analysis of Bioavailability and Bioequivalence Studies*. New York: Marcel Dekker.

Chow, S.-C, J. Shao, and H. Wang (2002). A note on sample size calculations for mean comparisons based on noncentral *t*-statistics. *Journal of Biopharmaceutical Statistics*, **12**, 441–456.

Chow, S.-C, J. Shao, and H. Wang (2008). *Sample Size Calculations in Clinical Research*, 2nd edition. Boca Raton, FL: Chapman and Hall/CRC.

Coffey, C. S. and K. E. Muller (1999). Exact test size and power of a Gaussian error linear model for an internal pilot study. *Statistics in Medicine*, **18**(1), 1199–1214.

Coffey, C. S. and K. E. Muller (2001). Controlling test size while gaining the benefits of an internal pilot study. *Biometrics*, **57**, 625–631.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Mahwah, NJ: Lawrence Erlbaum.

Cohen, Y. and J. Y. Cohen (2008). *Statistics and Data with R: An Applied Approach Through Examples*. Hoboken, NJ: Wiley.

Dattalo, P. (2008). *Determining Sample Size: Balancing Power, Precision, and Practicality*. New York: Oxford University Press.

Day, S. (2000). Operational difficulties with internal pilot studies to update sample size. *Drug Information Journal*, **34**, 461–468.

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *Journal of Statistical Planning and Inference*, **124**(1), 121–144.

De Santis, F. (2006). Sample size determination for robust Bayesian analysis. *Journal of the American Statistical Association*, **101**(473), 278–291.

De Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, **170**, 95–113.

Dendukuri, N., E. Rahme, P. Bélisle, and L. Joseph (2004). Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics*, **60**, 388–397. (Note that software is available for the methodology in the paper; see `http://www.medicine.mcgill.ca/epidemiology/joseph/Bayesian-Software-Bayesian-Sample-Size.html` and see on that page other software available corresponding to other papers.)

Denne, J. S. and C. Jennison (1999). Estimating the sample size for a *t*-test using an internal pilot. *Statistics in Medicine*, **18**(13), 1575–1585.

Diletti, E., D. Hauschke, and V. W. Steinijans (1992). Sample size determination for bioequivalence assessment by means of confidence intervals. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, **30**, 1–8.

Dixon, P. M. (1998). Assessing effect and no effect with equivalence tests.Chapter 12 in *Risk Assessment: Logic and Measurement* (M. C. Newman and C. L. Strojan, eds.). Ann Arbor, MI: Ann Arbor Press.

Dudewicz, E. J. (1972). Confidence intervals for power with special reference to medical trials. *Australian Journal of Statistics*, **14**, 211–216.

Faul, F., E. Erdfelder, A.-G. Lang, and A. Buchner (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, **39**(2), 175–191.

Fox, D. R., Y. Ben-Haim, K. R. Hayes, M. A. McCarthy, B. Wintle, and P. Dunstan (2007). An info-gap approach to power and sample size calculations. *EnvironMetrics*, **18**, 189–203.

Freiman, J. A., T. C. Chalmers, H. Smith, Jr., and R. R. Kuebler (1978). The importance of beta, the type II error and sample design in the design and interpretation of the randomized control trial: Survey of 71 “negative” trials. *New England Journal of Medicine*, **299**, 690–694.

Friede, T. and M. Kieser (2003). Blinded sample size reassessment in non-inferiority and equivalence trials. *Statistics in Medicine*, **22**, 995–1007.

Ganju, J., A. Izu, and A. Anemona (2008). Sample size for equivalence trials: A case study from a vaccine lot consistency trial. *Statistics in Medicine*, **27** (19), 3743–3754; Discussion, **28**, 175–179.

Garrett, K. A. (1997). Use of statistical tests of equivalence (bioequivalence tests) in plant pathology (Letter to the Editor). *Phytopathology*, **87**(4), 372–374.

Gauch, H. G. Jr. (2006). Winning the accuracy game. *American Scientist*, **94**(2), 133–141.

Glick, H. A. (2011a). Sample size and power for cost-effectiveness analysis (part 1). *Pharmacoeconomics*, **29**(3), 189–198.

Glick, H. A. (201lb). Sample size and power for cost-effectiveness analysis (part 2): The effect of maximum willingness to pay. *Pharmacoeconomics*, **29**(4), 287–296.

Goldstein, M. (1981). A Bayesian criterion for sample size. *Annals of Statistics*, **9**, 670–672.

Greenland, S. (1987). On sample-size and power calculations for studies using confidence intervals. *American Journal of Epidemiology*, **128**(1), 231–237.

Guenther, W. C. (1965). *Concepts of Statistical Inference*. New York: McGraw-Hill.

Guo, J. H., H. J. Chen, and W. M. Luh (2011). Sample size planning with the cost constraint for testing superiority and equivalence of two independent groups. *The British Journal of Mathematical and Statistical Psychology*, **64**(3), 439–461.

Gustafson, P. (2006). Sample size implications when biases are modelled rather than ignored. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, **169**, 865–881.

Halpern, S. D., J. H. T. Karlawish, and J. A. Berlin (2002). The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association*, **288**, 358–362.

Halpern, S. D., J. H. T. Karlawish, and J. A. Berlin (2005). Comment on "Ethics, and sample size" by Bacchetti, Wolf, Segal, et al. *American Journal of Epidemiology*, **162**, 195–196.

Hare, L. B. (2008). Statistics Roundtable: There is no such thing as parity. *Quality Progress* (January), 78–79.

Inoue L. Y. T., D. A. Berry, and G. Parmigiani (2005). Relationship between Bayesian and frequentist sample size determination. *The American Statistician*, **59**(1), 79–87.

Jiroutek, M.R., K. E. Muller, L. L. Kupper, and P. W. Stewart (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics*, **59**, 580–590.

Joseph, L. and P. Bélisle (1997). Bayesian sample size determination for normal means and differences between normal means. *The Statistician*, **46**(2), 209–226.

Joseph, L., R. du Berger, and P. Bélisle (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, **16**, 769–781.

Julious, S. A. (2010). *Sample Sizes for Clinical Trials*. Boca Raton, FL: Chapman and Hall/CRC.

Kieser, M. and T. Friede (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine*, **19**, 901–911.

Knottnerus, J. A. and L. M. Bouter (2001). The ethics of sample size: Two-sided testing and one-sided thinking. *Journal of Clinical Epidemiology*, **54**, 109–110.

Kraemer, H. C, J. Mintz, A. Noda, J. Tinklenberg, and Y. A. Yesavage (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, **63**, 484–489.

Kupper, L. L. and K. B. Hafner (1989). How appropriate are popular sample size formulas? *The American Statistician*, **43**(2), 101–105.

Lancaster, G. A., S. Dodd, and P. R. Williamson (2004). Design and analysis of pilot studies: Recommendations for good practice. *Journal of Evaluation in Clinical Practice*, **10**(2), 307–312.

Lautoche, A. and R. Porcher (2007). Sample size calculations in the presence of competing risks. *Statistics in Medicine*, **26**, 5370–5380.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, **55**(3), 187–193.

Lenth, R. V. (2006–2009). Java applets for power and sample size (computer software). (Available at `http://www.cs.uiowa.edu/~rlenth/Power`.)

Lindley, D. V. (1997). The choice of sample size. *The Statistician*, **46**(2), 129–138.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. North Scituate, MA: Duxbury.

Machin, D. and M. J. Campbell (2005). *The Design of Studies for Medical Research*. Hoboken, NJ: Wiley.

Matthews, J. N. S. (1995). Small clinical trials—are they all bad? *Statistics in Medicine*, **14**, 115–126.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, **9**(2), 147–163.

Maxwell, S. E., K. Kelley, and J. R. Rausch (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, **59**, 537–563.

Moyé, L. A. and A. T. N. Tita (2002). Defending the rational for the two-tailed test in clinical research. *Circulation*, **105**, 3062–3065. (Available at `http://circ.ahajournals.org/cgi/content/full/`105/25/3062.)

Noordzij, M., F. W. Dekker, C. Zoccali, and K. J. Jager (2011). Sample size calculations. *Nephron Clinical Practice*, **118**, 319–323.

O'Quigley, J. and C. Baudoin (1988). General approaches to the problem of bioequivalence. *The Statistician*, **37**, 51–58.

Parker, R. A. and N. G. Berman (2003). Sample size: More than calculations. *The American Statistician*, **57**(3), 166–170.

Pezeshk, H. (2003). Bayesian techniques for sample size determination in clinical trials: A review. *Statistical Methods in Medical Research*, **12**, 489–504.

Pezeshk, H, N. Nematollahi, V. Maroufy, and J. Gittins (2009). The choice of sample size: A mixed Bayesian/frequentist approach. *Statistical Methods in Medical Research*, **18**, 183–194.

Pham-Gia, T. and A. Bekker (2005). Sample size determination using Bayesian decision criteria under absolute value loss function. *American Journal of Mathematical and Management Sciences*, **25**(3/4), 259–291.

Phillips, K. F. (1990). Power of two-one-sided tests procedure in bioequivalence.*Journal of Pharmacokinetics and Biopharmaceutics*, **18**(2), 137–144.

Pocock, S. J. (2003). The pros and cons of noninferiority trials. *Fundamental and Clinical Pharmacology*, **17**, 483–490.

Posch, M. and P. Bauer (2000). Interim analysis and sample size reassessment. *Biometrics*, **56**, 1170–1176.

Rahardja, D. and Y. D. Zhao (2009). Unified sample size calculations using the competing probability. *Statistics in Biopharmaceutical Research*, **1**(3), 323–327.

Ryan, T. P. (2011). *Statistical Methods for Quality Improvement*, 3rd edition. Hoboken, NJ: Wiley.

Sahu, S. K. and T. M. F. Smith (2006). A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, **169**, 235–253.

Schulz, K. F. and D. A. Grimes (2005). Sample size calculations in randomised trials: Mandatory and mystical. *Lancet*, **365**, 1348–1353.

Schulz, K., D. Moher, and D. G. Altman (2010). A fundamental misinterpretation of CONSORT. Comment on Bacchetti (2010). *BMC Medicine*, **8** (online journal; `http://www.biomedcentral.eom/1741-7015/8/17/comments#414700`).

Sedlmeier, P. and G. Gigerenzer (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, **105**, 309–316.

Shao, J. and S.-C. Chow (2002). Reproducibility probability in clinical trials.*Statistics in Medicine*, **21**(12), 1727–1742.

Shih, W. J. (2001). Sample size re-estimation — journey for a decade. *Statistics in Medicine*, **20**, 515–518.

Simon, S. (2008). Too much power and precision? (Web page: `http://www.pmean.com/08/TooMuchPower.html`.)

Sims, M., D. A. Elston, M. P. Harris, and S. Wanless (2007). Incorporating variance uncertainty into a power analysis of monitoring designs. *Journal of Agricultural, Biological, and Environmental Statistics*, **12**(2), 236–249.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, **24**, 243–258.

Strug, L. J., C. A. Rohde, and P. N. Corey (2007). An introduction to evidential sample size calculations. *The American Statistician*, **61**, 207–212.

Stüger, H. P. (2006). Asymmetric loss functions and sample size determination: A Bayesian approach. *Austrian Journal of Statistics*, **35**, 57–66.

Sutton, A. J., N. J. Cooper, D. R. Jones, P. C. Lambert, J. R. Thompson, and K. R. Abrams (2007). Evidence-based sample size calculations based upon updated meta-analysis. *Statistics in Medicine*, **26**, 2479–2500.

Taylor, D. J. and K. E. Muller (1995). Computing confidence bounds for power and sample size of the general linear model. *The American Statistician*, **49**(1), 43–47.

Thomas, S. (1999). Retrospective power analysis. *Conservation Biology*, **11**, 276–280.

van Belle, G. (2008). *Statistical Rules of Thumb*, 2nd edition. Hoboken, NJ: Wiley.

Vickens, A. J. (2003). Underpowering in randomized trials reporting a sample size calculation. *Journal of Clinical Epidemiology*, **56**, 719–720.

Walker, S. G. (2003). How many samples? A Bayesian nonparametric approach. *The Statistician*, **52**, 475–482.

Wang, F. and A. E. Gelfand (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, **17**(2), 193–208.

Wang, J. and J. D. Stamey (2010). A Bayesian algorithm for sample size determination for equivalence and non-inferiority test. *Journal of Applied Statistics*, **37**, 1749–1759.

Watson, W. (2008). Updates to SAS power and sample size software in SAS/STAT 9.2. Paper 368-2008.

Webb, R. Y., P. J. Smith and A. Firag (2010). On the probability of improved accuracy with increased sample size. *The American Statistician*, **64**, 257–262.

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician*, **46**(2), 185–191.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd editon. Boca Raton, FL: CRC Press.

Wilburn, A. J. (1984). *Practical Statistical Sampling for Auditors*. Boca Raton, FL: CRC Press.

Wittes, J. and E. Brittain (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, **9**, 65–72.

Wittes, J., O. Schabenberger, D. Zucker, E. Brittain, and M. Proschan (1999). Internal pilot studies I: Type 1 error rate of the naive *t*-test. *Statistics in Medicine*, **18**(24), 3481–3491.

Yin, K., P. K. Choudhary, D. Varghese, and S. R. Goodman (2008). A Bayesian approach for sample size determination in method comparison studies. *Statistics in Medicine*, **27**, 2273–2289.

Zhang, P. (2003). A simple formula for sample size determination in equivalence studies. *Journal of Biopharmaceutical Statistics*, **13**(3), 529–538.

Zhao, W. and A. X. Li (2011). Estimating sample size through simulations. PharmaSUG2011, Paper SP08.

Zucker, D. M., J. T. Wittes, O. Schabenberger, and E. Brittain (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine*, **18**, 3493–3509.

EXERCISES

2.1. Assume a $N(\mu, \sigma^2)$ distribution, with $\mu = \sigma^2 = 16$. Solve for *n* such that the maximum possible error of estimation of μ is to be 2 with probability .95. Then generate 10,000 samples of size *n,* with the value of *n* being what you solved for. What percentage of observations is within 2 units of μ? (Obviously a computer program will have to be written to generate the 10,000 samples and to determine the percentage.) What have you demonstrated?

2.2. Explain how a study could have "too much power." How could this be prevented?

2.3. The following request was (tersely) made on a Web blog some years ago: "Alpha .05, Power 0.8. What is the sample size to detect an outcome difference of .20 versus .30 for an adverse event. Thank you." This was used as a lead-in to the author's mention of websites that will do this type of calculation. If you were a consultant and you received such a request, would you refer the person asking the question to a website or two, or would you react differently? Explain.

2.4. As explained in Section 2.2, there is no reason why designed studies should always focus on a power of .80. There is some dialogue from an episode of the television show *Walker, Texas Ranger* that can be used to dramatize this point. A mobster will soon be on trial and his attorney states: "I've run some simulations and there is an 83% chance that you will be acquitted. I must say those are excellent odds." The mobster then replies: "Oh, really. That's like 5 out of 6, isn't it?" The attorney replies "Yes sir." The mobster then puts a bullet in a gun, points the gun at the attorney and states: "One bullet, one chamber. If I spin the chamber and fire the gun, there is an 83% chance that I won't blow you away. Do you still think those are excellent odds?" The attorney then replies "No sir." Similarly, if you are conducting a study whose results might greatly benefit humankind, do you want to have an 80% chance of making such an important discovery, or would you want the odds to be more in your favor?

2.5. Assume that a study was performed and approximate normality was a reasonable assumption for the population of observations. The standard deviation was unknown but values below 10 or above 70 are rarely observed. Without giving any regard to power, an experimenter insists on using 100 observations and is interested in detecting a five-unit increase in the population mean with probability .90, using a one-tailed test with $\alpha = .05$. What was the power of the test? Would you recommend that the experimenter use a larger or a smaller sample size in a future study involving the same population? Explain your recommendation.

2.6. Explain why the assumed power for a study will almost certainly not be the actual power.

2.7. An experimenter wants to determine the sample size for a 95% confidence interval for μ that will be used to test the null hypothesis that $\mu = 50$ and have a probability of .95 of rejecting the null hypothesis when the true mean exceeds 50 by one standard deviation of the mean. Can this be accomplished? Explain.

2.8. Often the planned sample size will not be the actual sample size because circumstances may prevent the number of subjects being available that was originally proposed (people can drop out of a study, data can be lost, etc.). Assume that the sample size of 75 was determined to provide a power of .90 but a sample size of only 68 subjects could be used. Does this automatically mean that the *actual* power of the study was less than .90? Explain.

2.9. If some of the participants in a study subsequently withdraw, will the power of the study increase or decrease? Explain.

2.10. If, after a study has been initiated, a scientist decides that it is more important to detect a smaller effect than the effect stated when the sample size for the study was computed, if that sample size is used and everything else remains the same, will the power for detecting this smaller effect be larger or smaller than the power that was originally specified? Explain.

2.11. Assume that a sample size was determined for an upper one-tailed test but there was a miscommunication and the test to be used will actually be a lower one-sided test, with the difference to be detected from the hypothesized mean being the same in absolute value as it would have been if the upper one-tailed test had been used. If nothing else changes, will it be necessary to recompute the sample size or will it be the same? Explain.