CHAPTER 3

Means and Variances

In this chapter we consider power and sample size determination for a mean for a single population and means for two populations, for both independent and dependent samples, and also cover sample size for testing a single variance and for testing the equality of two variances. We also briefly discuss the case of more than two means, with this covered fully in Chapter 6 in the context of experimental designs.

The emphasis in this chapter and in subsequent chapters is on hypothesis testing because that is what is emphasized in the literature and in software. Sample size determination for confidence intervals is also covered since it is also important and has been covered in the literature in articles such as those by Kelley and Rausch (2006), Jiroutek, Muller, Kupper, and Stewart (2003), Bristol (1989), and Beal (1989). See also Grieve (1991), who examined the suggestion of Beal (1989).

It is highly desirable for researchers to indicate how they obtained the sample sizes that they used in their studies because various assumptions must be made in the calculation of sample sizes, including parameter values, which will almost always be unknown. Nevertheless, Nayak (2010) cited the study of Moher, Dulberg, and Wells (1994), who found that sample size calculations were given in only 32% of studies that did *not* result in statistical significance. This could have occurred in some studies because the study was underpowered and/or unrealistic assumptions were made, such as bad inputs for parameter values. The reader of research articles needs enough information to be able to determine if the study was not well designed, so both the sample size and the manner in which it was determined should be provided in research articles.

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

3.1 ONE MEAN, NORMALITY, AND KNOWN STANDARD DEVIATION

Example 2.1 was used partly to show how one could obtain an initial estimate of σ, which is necessary for sample size determination and power computation, without obtaining the estimate from sample data, such as in a pilot study, although the desirability of using an internal pilot study was emphasized in Chapter 2. (Note that software is available to aid in the planning of pilot studies, including Lenth's applet, as was discussed and illustrated in Section 2.1.)

In this section, we discuss designing a study for testing a single mean in more detail than was given in Example 2.1. The expression for n was given in that example for a one-sided hypothesis test. The general expression given in Eq. (2.3) for a one-sided test is

$$n = \left[\frac{(Z_\alpha + Z_\beta)\sigma}{\mu - \mu_0} \right]^2 \tag{3.1}$$

which, among other things, shows that a sample size necessary to detect a small departure from μ_0 requires, for fixed α, β, and σ, a larger sample size than when the departure is not small. For a fixed difference,$\mu - \mu_0$, with μ_0 denoting the hypothesized mean and μ denoting the mean value that one wishes to detect with the stated power, the sample size is increased if either α or β is decreased, or if σ is increased. This should be intuitive because decreasing β means that the power, $1 - \beta$, is increased, and a more powerful test requires, other things being equal, an increased sample size. Although the power or n can be changed and the other one will also change in the same direction, α is chosen independently of n and β. In "modern" hypothesis testing, α might not be specified before the data are collected and analyzed; instead, a decision would be reached based on the magnitude of the p-value. A benchmark value for the p-value must be used, however, because the experimenter must decide if the p-value is small enough to reject the null hypothesis. That decision is certainly going to depend on the field of application and the type of study that has been conducted.

In general, the smaller the value of α, the less area under the curve for the actual distribution will lie in the rejection region, as can be seen from Figure 3.1.

Therefore, decreasing α (i.e., increasing Z_α) will decrease power (increase β, which will decrease Z_β). These changes will be somewhat offsetting in Eq. (3.1). Consequently, sample size is most straightforwardly viewed as being most strongly influenced by the difference between the parameter value that an experimenter wishes to detect and the hypothesized parameter value. (Of course, σ will also have an effect on the sample size and it will also strongly affect the power for a given sample size if a bad estimate of σ is used as input.)

The expression for n is slightly more involved when a two-sided test is used, which would be used if the experimenter simply wanted to detect a difference

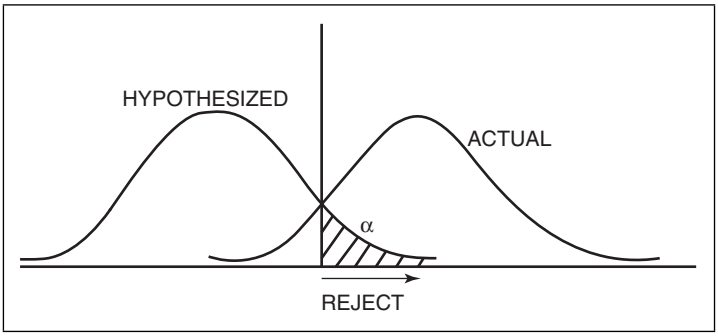

Figure 3.1 Hypothesized distribution and actual distribution.

from the hypothesized parameter value, without having a preconceived notion of the direction of the departure.

As an approximation that should usually work reasonably well (which of course would be unnecessary if software were used), we could simply replace Z_α by $Z_{\alpha/2}$ in Eq. (3.1). If we did so, we would be acting as if the test was actually a one-sided test using a significance level of $\alpha/2$ rather than a two-sided test with significance level α. Although this might seem improper, we need to take a commonsense approach to determining sample size and computing power for a two-sided test. Specifically, if the true value of a population parameter is greater than the hypothesized value, should the information about the assumed distribution on the side opposite the true value have any influence on sample size determination and power? For example, assume that $\mu_0 = 50$ and $\mu = 51$. If there is a small difference between the hypothesized value and the true value, relative to σ, a value of the sample mean below 50 could result, and the value of the test statistic could be negative and fall in the rejection region on the low (opposite) end. Should that possibility influence sample size determination? It should probably not because if, say, the sample mean were 48 and the value of the test statistic was in the rejection region, we would not logically conclude that the mean was some value in excess of 50. We would simply conclude that the mean was not 50. When we calculate power, however, we can't calculate it for something like "not 50"; we have to specify a value. Therefore, one could argue that it is logical to simply replace Z_α by $Z_{\alpha/2}$ in Eq. (3.1) and use that expression to solve for the sample size. We will see later that this will often give the correct result, although it is not necessarily recommended as a routine procedure. Of course, we should also keep in mind that we are "approximating" anyway because σ is never known. (See the discussion in Section 3.2 of a way to address the problem of σ being unknown.)

Note: For the examples that follow in this chapter and throughout the remainder of the book, it is important to recognize the difference between *assumed power* and *actual power*, as well as *target power*, as used by some software. The

assumed power is what the power would be (i.e., the actual power) if the true values for all "nuisance parameters" (such as sigma when sample size is being determined for testing a population mean) are equal to the values that are entered into software. Since this will hardly ever happen, we should think of the assumed power as almost always incorrect and not equal to the actual power. Certain software, such as MINITAB, will show "target power" and "actual power" in juxtaposition in output when a discrete random variable is involved, as then there is a difference between target power and assumed power since target power cannot be achieved, in general, even if nuisance parameter values were known, because of jumps in cumulative probabilities between successive possible values of the random variable. Thus, for example, the assumed power might jump from .884 to .918 and bypass .900. Such output labeling is actually incorrect because what is being called actual power in the output is really retrospective power, and as explained in Section 1.4, it doesn't make any sense to use that term.

■ EXAMPLE 3.1

To illustrate the use of Eq. (3.1) and its modification, we will assume $\alpha = .05$, power $= .80$, $\sigma = 3$, $\mu_0 = 50$, and $\mu = 52$. For a one-sided test, Eq. (3.1) applies directly and we obtain $n = 13.91$, so $n = 14$ would be used. When software such as PASS, Power and Precision, or MINITAB is used, $n = 14$ results with actual power of .802.

As implied in Section 2.5 and by Parker and Berman (2003), the user may be interested in more than a specific difference $(\mu - \mu_0)$ and the corresponding power, so the power curve in Figure 3.2, produced by MINITAB, can be useful.

We generally won't be able to hit the desired power exactly simply because sample size is, of course, discrete, so there will be jumps in the power value per unit change in the sample size, and the power will not generally be a round number. Here the power is slightly greater than .80 because 14 is slightly greater than 13.91. Curiously, Power and Precision gives the two-sided confidence interval when the user specifies that sample size is to be determined for a one-sided hypothesis test. (The confidence interval limits are computed using the entered value of σ and value of μ specified in the alternative hypothesis. That is, assumed parameter values are used, which is not the way that orthodox confidence intervals are constructed.)

For a two-sided test, substituting $Z_{\alpha/2}$ for Z_α in Eq. (3.1) results in $n = 17.66$. When software such as MINITAB or Power and Precision is used, $n = 18$ results with actual power of .807. The corresponding power curve then contains the portion shown in Figure 3.2 plus the mirror image of that curve, with the full curve symmetric about 0. ■

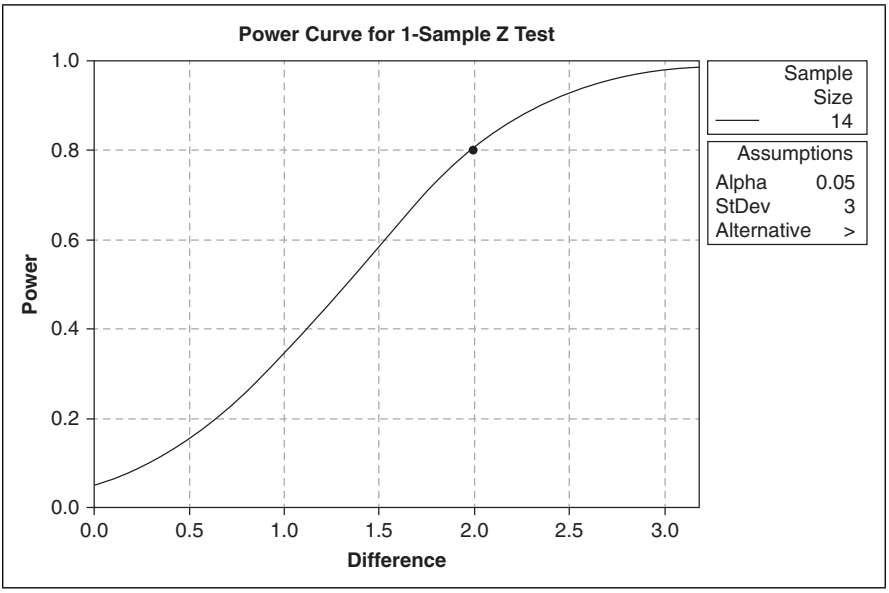

Figure 3.2 Power curve.

It was stated in Section 3.1 that using either .95 or .99 for power could, depending on the value of the standard error, require such a large sample size as to be impractical. For example, if the desired power in Example 3.1 had been .95 or .99 instead of .80, the necessary sample size would have been 25 for power of .95 (actually .9543), and 36 for power of .99. Although the latter sample size is not large, it is almost three times the sample size that is needed for a power of .80.

Although there is agreement between the use of

$$n = \left[\frac{(Z_{\alpha/2} + Z_{\beta})\sigma}{\mu - \mu_0}\right]^2 \tag{3.2}$$

for a two-sided test and the result obtained using software for this example, that won't always be the case. We can see what is happening if we consider the expression that is used as the starting point in solving for n in the two-sided case:

$$1 - \Phi\left(Z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) + \Phi\left(-Z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) = 1 - \beta \tag{3.3}$$

which is derived in the chapter Appendix. The first two terms on the left side are the same as those used in deriving the sample size for an upper one-sided test, except that Z_{α} is used instead of $Z_{\alpha/2}$ (with the third term, using Z_{α}, used for a

lower one-sided test), so if the third term is essentially zero for the computed sample size, then the third term is not having any effect on the determination of the sample size. If the third term is not very close to zero, however, as could happen if μ_0 was greater than μ, and the difference was small such that the sum of the first two terms in Eq. (3.3) is not approximately zero, then Eq. (3.2) will only approximate the sample size. Since a closed-form expression for n cannot be obtained because Φ^{-1} is not a linear operator, the use of software is highly desirable. [Of course, Eq. (3.3) is easy to solve numerically for n.]

In general, Eq. (3.2) would give a sample size that is off slightly whenever $(\mu_0 - \mu)/(\sigma/\sqrt{n})$ is small because then neither the third term nor the sum of the first two terms in Eq. (3.1) would be close to zero. Fortunately, however, the fraction really can't be very small because a large value of n is needed to detect a small difference $(\mu_0 - \mu)$, and if n is large, the denominator of the fraction will be small. Of course, the fraction could be small if σ is quite large, but we would expect the difference $(\mu_0 - \mu)$ to be related to σ, such as perhaps being equal to σ. So these effects would be offsetting.

To illustrate, if $\mu = 50.8$, Eq. (3.2) gives $n = 111$, which agrees with the result obtained using software including MINITAB, Power and Precision, and PASS if we select the "variance known" option of Power and Precision and the "known standard deviation" option of PASS so that a z-test is assumed, with such specification unnecessary using MINITAB since the 1-sample Z routine assumes a known standard deviation. (Note that such an option is not available with Lenth's applet, as it assumes the use of a t-test. This results in $n = 113$, so the difference is only slight, and it can be shown that the difference is also slight, and virtually constant, for similar values of $|\mu_0 - \mu|$.)

Similarly, if $\mu = 50.6$, $n = 197$ by Eq. (3.2) and by MINITAB, PASS, and Power and Precision. In the first case, $(\mu_0 - \mu)/\left(\sigma/\sqrt{n}\right) = 2.80951$ (using $n = 111$) and 2.821 in the second case—almost the same. Even when $\mu = 50.2$, so that $n = 1766$, the ratio is 2.802, which differs very little from the other two values. Of course, we would expect the fraction to be essentially constant at about 2.8 because the first term is essentially 1.0 when the fraction is about 2.8, since $\Phi(4.76) \approx 1$ and thus

$$1 - \Phi\left(Z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right)$$

is not going to make a contribution to the desired power value of .80, and $\Phi(-1.96 + 2.8) = \Phi(.84) = .80$.

In essence, a huge sample size shrinks the distribution of $\bar{X}$ so much that there is hardly any spread of values. The area under the curve for the distribution with assumed mean = 50.1 that is in the nonrejection region for the distribution with the hypothesized mean is .20, as can be seen in Figure 3.3. Therefore, the power is .80.

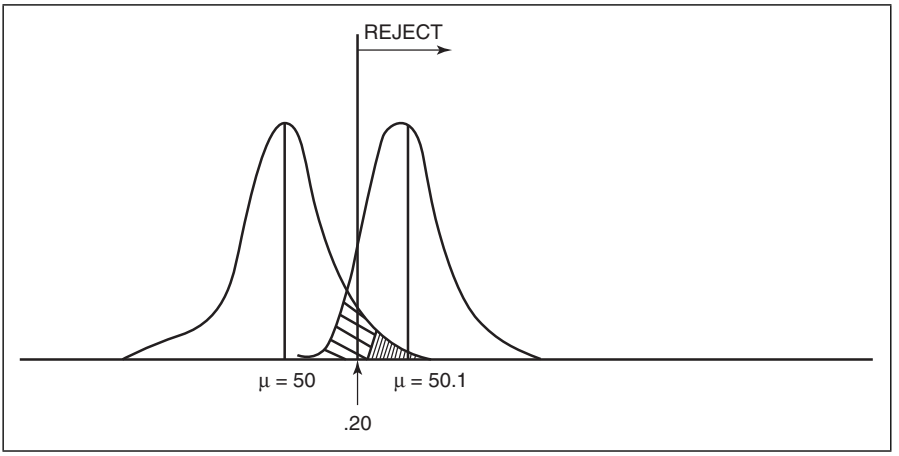

Figure 3.3 Effect of huge sample size.

Using software to obtain sample sizes renders unnecessary the need for simple approximations, unless a user does not have access to software or does not want to use formulas that might be viewed as complicated. Van Belle (2008, Chapter 2) gave a simplified formula for the one-sample case and a two-sided hypothesis test, which is $n = 8/\Delta^2$, with $\Delta = [(\mu - \mu_0)/\sigma]^2$, with μ denoting the true population mean and μ_0 denoting the hypothesized mean value. This will provide a reasonable approximation only if $\alpha = .05$ and power $= .80$, however, as then the two Z-values in Eq. (3.2) are 1.96 and 0.8416, respectively, the sum of which is 2.8016 and the square of that sum is 7.85—close to 8.0. Since a user would generally not be interested in detecting a small difference from the hypothesized value, Δ^2 probably won't be very small, so the approximation will usually be reasonably good. If a user preferred to have power $\geq .90$, however, the approximation would likely be poor since $(1.96 + 1.28)^2 = 10.50$ for power $=$ .90. Thus, the approximation could be applied only when the test is two-sided with $\alpha = .05$ and power $= .80$ for detecting the effect size Δ.

■ EXAMPLE 3.2

In illustrating applications in orthodontics, Newcombe (2001) focused attention on determining sample size so as to give a confidence interval for a population mean with a specified width. (Recall the discussion of this in Section 2.4.) The objective was to estimate the mean toothbrushing force before the attachment of an orthodontic appliance. Newcombe (2001) used as an estimate of σ the sample standard deviation from a similar study described by Heasman, MacGregor, Wilson, and Kelly (1998). That study, which had an objective of determining whether or not toothbrushing forces were affected by wearing fixed orthodontic

appliances, had a sample standard deviation of 124, with 30 children (10 males and 20 females, ages 10–15) used in the study.

There are some questions that should have been addressed regarding the use of that estimate of σ, one of which would seem to be the spread of the ages of children for the study in Newcombe (2001) relative to the spread of the ages of the children in the Heasman et al. (1998) study, as this would probably be related to the standard deviation. This issue was not addressed by Newcombe (2001), however. The use of a standard deviation from a study performed by different investigators is potentially shaky regarding the comparability of the two sets of subjects. When comparability cannot be assessed, it seems almost imperative to use the upper limit of a confidence interval for σ rather than using simply the standard deviation from the other study.

Indeed, using the upper limit of an 80% confidence interval has been proposed in the literature, as Browne (1995) showed by the use of simulations that using a $100(1-\alpha)\%$ upper one-sided confidence limit on σ will provide a sample size that is sufficient to achieve the desired power $100(1-\alpha)\%$ of the time. This idea is discussed further in Section 3.3. Kieser and Wassmer (1996) provided some analytical insights into Browne's proposal.

We can see how this would affect the sample size for the Newcombe (2001) study. No distributional assumption was made for the latter although approximate normality was tacitly assumed by the assumption that a 95% confidence interval for the mean is approximately $\bar{x} \pm 2\hat{\sigma}/\sqrt{n}$ when n is large. The initial objective was to select n so that $2\hat{\sigma}/\sqrt{n} = 30$ (i.e., the confidence interval width would be 60). Solving for n produces $n = (2\hat{\sigma}/30)^2 = [2(124)/30]^2 = 68.34$, so $n = 68$ or 69 could be used. Newcombe (2001) used $n = 68$ and since the multiplier "2" is being used as an approximation (instead of, say, 1.96), there is not strong motivation for rounding up, as is done in hypothesis testing.

If we assume normality, following Browne (1995), we could use at least an 80% one-sided upper confidence bound for σ. Since a standard deviation is being used from a study performed by different investigators and the two sets of subjects may differ in important ways, a 95% upper confidence bound may be preferable. The upper bound is obtained as

$$\hat{\sigma}\sqrt{\frac{n-1}{\chi^2_{1-\alpha,\,n-1}}} = 124\sqrt{\frac{29}{17.7084}} = 158.68$$

Using the latter in place of 124 produces a sample size of $[2(158.68)/30]^2 = 111.9$, so $n = 112$ would be used. This would be a much safer, albeit also more costly, sample size.

Newcombe (2001) stated that it might be decided that a confidence interval of $\bar{x} \pm 30$ could be deemed too wide to be informative. If 30 is replaced by 15, however, the required sample size will be four times the original sample size, which we can see without doing any calculations since $15 = 30/2$ and the 2 will

be squared in performing the computation. Newcombe (2001) indicated that this would require a sample size of 273, whereas using the upper confidence bound for σ would result in a sample size of 448. Both 448 and 273 may be impractical or even impossible for the new study, however. At the very least, the cost would be greatly increased.

Could software be used for this problem to determine sample size? The answer is "yes" although slightly fewer options exist since software for sample size determination emphasizes hypothesis tests with less attention devoted to confidence intervals. PASS is one software that can be used for this problem and it does have moderately extensive capability for confidence intervals. The user would select "Confidence intervals" from the main menu and then select "Means." It is necessary to enter .9544 for the confidence level since that corresponds to the 2σ limits used by Newcombe (2001). Doing so and entering 30 for the "Distance from Mean to Limits," then checking "Known Standard Deviation," indicating a two-sided interval and entering 124 for σ results in a sample size of 69, in general agreement with the hand computation result of 68.34. Similarly, using 158.68 as the upper bound estimate of σ results in $n = 112$, in agreement with the hand computation result of 111.9. These results are also obtained using nQuery, which has overall confidence interval capabilities similar to that of PASS.

These results can also be produced using MINITAB (Release 16 or later). Specifically, the user selects "Power and Sample Size" from the main menu, then selects "Sample Size for Estimation" and then "Mean (Normal)" from the options that are available. Then entering 124 for the standard deviation, 30 for the margin of error, indicating that the standard deviation is known, and specifying 95.44 for the degree of confidence produces $n = 69$. Similarly, when 158.68 is used for the assumed known value of σ, $n = 112$ is produced, with both of these results in agreement with the results given by PASS and nQuery. Finally, with Lenth's software it is not possible to specify a 95.44% confidence interval so the user would have to settle for a 95% confidence interval. Entering the same inputs as when the other software was used produces $n = 68.06$ and $n = 109.9$, respectively, for the two inputted values of σ. Thus, there are some small differences, especially for the larger inputted value of σ, which is due to the fact that the degree of confidence is not matched exactly. ■

3.1.1 Using the Coefficient of Variation

Van Belle and Martin (1993) pointed out that the information provided by a consulting client will often be in terms of the percentage change expected from a treatment, the percentage change that would be clinically significant, and the "percentage variability" in the data. Of course, the latter is rather vague but the authors stated that this can be a lead-in to estimating a coefficient of variation (*CV*). (The population *CV* is defined as μ/σ and the sample *CV* defined as $\bar{X}/s$.) Furthermore, the client may be interested in a percentage change in the mean.

Van Belle (2008, p. 34) pointed out that we can define a percentage change using either the current mean or the mean to be detected in the denominator, which of course is obvious. Van Belle uses the average of the two means in the denominator, however, and defines the percentage change, *PC*, as

$$PC = \frac{\mu_0 - \mu_1}{\mu}$$

and then, apparently assuming $\alpha = .05$ and power $= .80$, states that the sample size is "estimated remarkably accurately" by

$$n = \frac{16(CV)^2}{(PC)^2}$$

Since a coefficient of variation of 35% is apparently quite common in biological systems (van Belle, 2008, p. 34), a rule-of-thumb for the sample size in such applications is then

$$n = \frac{2}{(PC)^2}$$

since $16(0.35)^2 = 1.96$.

I am not aware of any software that can be used for sample size determination when the value of a coefficient of variation is used as input, so hand computation is apparently necessary. As should be apparent, the computation is quite simple, especially if the van Belle rule-of-thumb is used. It should be kept in mind, however, that this is indeed just a simple rule-of-thumb.

3.2 ONE MEAN, STANDARD DEVIATION UNKNOWN, NORMALITY ASSUMED

When σ is unknown (the usual scenario), the sample size formula for n is, as we logically expect, obtained by substituting $t_{\alpha, n-1}$ for Z_α in Eq. (3.1) and substituting $t_{\beta, n-1}$ for Z_β. Of course, this does not produce a simple expression that can be used to solve for n directly since the two t-variates are each a function of n. Thus, n must be solved for using iteration. We can, however, assuming a one-sided test and $\hat{\sigma}$ being some estimate of σ, use the expression

$$n = \left[\frac{\left(t_{\alpha, n-1} + t_{\beta, n-1}\right) \hat{\sigma}}{\mu - \mu_0} \right]^2 \tag{3.4}$$

to obtain the same value of n that is obtained using software, thus illustrating the formula computation.

To illustrate, assume that $\alpha = .05$, power $= .80$, $\hat{\sigma} = 3$, $\mu_0 = 50$, and $\mu = 52$, and the test is one-sided. If we assume that $\sigma = 3$ and use Eq. (3.1), we obtain $n = 13.91$, so $n = 14$ would be used. We would expect a slightly larger value of n to result from the assumption that σ is unknown. Although we can't use Eq. (3.4) to solve for n directly, we can use it to verify the solution obtained from software. That solution is $n = 16$, which gives a power value of .8156. Thus, when we use $n = 16$ in Eq. (3.4), we would expect to obtain a solution that is closer to 15 than 16 since the power for $n = 15$ is .7908. That is what happens as the computed value is 15.44.

Thus, the assumption of an unknown σ has resulted in a sample size that is two units larger. In both cases, however, we had to input a value of σ, thus showing the artificiality of the computation for the case of σ unknown. Early work on sample size determination for t-tests included Guenther (1981).

3.3 CONFIDENCE INTERVALS ON POWER AND/OR SAMPLE SIZE

Confidence intervals on power are not generally computed and are not available in sample size determination software. The potential usefulness of such an interval should be apparent, however, since it is important to recognize that power is rarely known because it partially depends on σ, which is unknown and must be estimated. Such estimation, however it is performed, causes the assumed power to actually be estimated power and to be a random variable if the estimate is obtained using historical data, which might be data from a pilot study.

For a one-sample t-test, we would expect, for a fixed sample size and significance level, the width of a confidence interval on power to be directly related to the width of a confidence interval on σ, if such an interval were constructed. Under the assumption of normality of the individual observations, a $100(1-\alpha)\%$ confidence interval for σ is given by

$$\text{Lower Limit: } s\sqrt{\frac{n-1}{\chi^2_{a/2,\,n-1}}} \qquad \text{Upper Limit: } s\sqrt{\frac{n-1}{\chi^2_{1-a/2,\,n-1}}}$$

(Note that subscripts on χ^2 are typically written, as is done here, such that $\alpha/2$ designates the cumulative area, not the area in the right tail.) Thus, the width depends on the sample size and the value of the estimate of σ, which in turn would logically be related to the magnitude of σ.

Dudewicz (1972) suggested substituting exact confidence bounds for $\hat{\sigma}$ into the power calculations for a t-test. This would have the effect of producing approximate confidence limits on the power of the test and would be more

realistic. Of course, this could easily be performed with software as the endpoints of the interval could be entered, in turn, in the *t*-test routine, fixing the value of n at perhaps the value that resulted from the use of a point estimate of σ.

The limits would be only approximate because the general form of the limits is based on the assumption that σ is known. Substituting a value for σ into an expression developed for σ known is not the same as using the appropriate expression constructed under the assumption that σ is unknown and is being estimated. Therefore, it is highly desirable to determine the worth of the approximate limits. Dudewicz (1972), however, did not, as indicated by Taylor and Muller (1995), note that this was only an approximate approach and thus did not provide any asymptotic or simulation results indicating the extent to which these limits can be expected to deviate from the correct limits, and apparently this has not been investigated to any extent.

Recognizing the need to numerically investigate the Dudewicz approach, Taylor and Muller (1995) used simulation to examine the method for one-way analysis of variance and found that the results support use of the method, which were in general agreement with the asymptotic result given by Clark (1987) that the method provides asymptotically unbiased confidence intervals for power. They recommended that a one-sided (lower) confidence bound be used for power since there is generally interest in having power at least equal to a target value. They also recommended an upper confidence bound on the sample size and gave the methodology for obtaining the upper bound, which requires the iterative solution of two equations. They also proposed the use of simultaneous confidence bounds for power and sample size and suggested the use of graphs that show exact confidence limits on power.

■ EXAMPLE 3.3

Taylor and Muller (1995) illustrated their methodology by applying it to a research problem involving deteriorating renal function that was addressed by Falk, Hogan, Muller, and Jeannette (1992). The latter defined a clinically significant improvement in renal function as a doubling of reciprocal serum creatinine level. Thirteen patients were randomly assigned to each treatment in a clinical trial and the use of a variance estimate of 0.68 resulted in a power estimate of .92. Taylor and Muller (1995) addressed the uncertainty in that number and obtained a 95% confidence interval on power as [.688, .999]. Of course, that is a very wide interval, with the upper limit being overkill and the lower limit being a value that most researchers would probably consider unacceptable. They stated that increasing the number of patients assigned to each treatment from 13 to 17.95 (i.e., 18) would ensure with probability .975 a lower bound on power of .90. Thus, for this study, the actual power may have been much lower than the assumed power, so an increase in the number of patients assigned to each group would have been desirable. If there was

a considerable per-patient cost involved, the researchers might have settled for 16 or 17 patients per group since they might have been willing to accept a lower bound on power of slightly less than .90, but the need for providing researchers with such a confidence bound should be obvious. ■

In related and more recent work, Wong (2010) explained how to obtain a confidence interval on the power of the one-sample *t*-test, with the alternative hypothesis presumed to be $\mu = \mu_0 + k\sigma$, with μ_0 denoting the mean under the null hypothesis. Sample size could also be determined as a by-product, but this would require an iterative solution. Consequently, it is not likely to be used by practitioners unless it is incorporated into the software most frequently used for sample size determination. (The authors did state that R code is available on request.)

As stated by Wong (2010), Lehmann (1959) proved that the probability of committing a Type II error for a one-sided, one-sample *t*-test with significance level α is given by

$$\beta = G_{n-1,\, k\sqrt{n}}(t_{n-1,1-\alpha}) \tag{3.5}$$

with $G_{\nu,\, a}\ (\cdot)$ denoting the cumulative distribution function of the noncentral *t*-distribution with ν degrees of freedom, noncentrality parameter, a, and $k = (\mu - \mu_0)/\sigma$. Of course, n is unknown but could be solved for numerically using Eq. (3.5) for a selected value of power $= 1 - \beta$ and thus β.

Certainly σ is also generally unknown and would have to be estimated unless one wishes to bypass that issue by specifying a value for k, which represents the effect size. If that approach were taken, a point estimate of power and a confidence interval for it could be computed before data are obtained, thus avoiding the type of criticism that has been leveled at retrospective power, but incurring the type of criticism that Lenth (2001) stated for dealing directly with effect sizes.

Deviating from the methodology given by Wong (2010), let's assume that this was not done, but rather that σ was estimated using prior information and *then* the effect size was computed. Similarly, we will specify upper and lower bounds on σ (without data), and we will need to specify a tentative sample size for the confidence interval on power. (Of course, σ might be estimated from data in a small pilot study; this will be illustrated later.)

Unfortunately, such methods are not available in sample size determination software and it seems unlikely that they will be available in the foreseeable future.

In general, it would be desirable for the methodology advocated by Dudewicz (1972) to be extended to other types of tests, with simulation results, preferably, or at least asymptotic results indicating how well the method performs. This has apparently not been discussed in the literature, however, but the extension would be reasonably straightforward when only a single parameter had to be estimated, such as in simple linear regression when testing the slope parameter,

as a confidence interval for σ^2_{error} would be needed, although a value for the spread of the regressor values would also be required. At the other extreme, it would be very complex and perhaps even intractable when confidence bounds (or a joint confidence region) must be developed for handling multiple unknown parameters, as when an entire variance–covariance matrix must be estimated. Consequently, what some discerning readers might call a naive approach of substituting values for unknown parameters will continue to be used throughout this book because that is simply all that is available—both in the literature and in software, although of course software could be used with confidence limits on σ^2 simply by using the appropriate routine twice, using the lower limit and then the upper limit on σ^2.

If there is any faith at all in the value of σ that is inputted for use with the t-statistic, then the t-statistic should not necessarily be used, as that utilizes the sample standard deviation, computed after the sample has been taken. Assuming normality, the sample standard deviation is biased, with the amount of bias a function of the sample size, as is Var(s). If E(s) is approximately equal to the value of σ that is inputted and software gives a large value of n, then using the t-statistic approach is not particularly bad, but we would nevertheless generally use a z-statistic if we strongly believe that we have an excellent estimate of σ that is inputted. Remember that there is very little difference between the values of t-variates and the corresponding z-variates for large sample sizes.

3.4 ONE MEAN, STANDARD DEVIATION UNKNOWN, NONNORMALITY ASSUMED

The t-test for a single mean is not undermined by slight-to-moderate nonnormality. However, when there is considerable nonnormality, a t-test should not be used. Assumptions are tested with data, but if an experimenter is at the stage of trying to determine the sample size, the data have not yet been obtained. So there is a problem unless data are available from a prior study. Subject matter knowledge might at least suggest the general shape of the distribution, and that knowledge might be used along with a book on distributions that shows the shape of each distribution for various combinations of parameter values. Evans, Hastings, Peacock, and Forbes (2010) is one such book.

For example, a t-distribution has heavier tails than a normal distribution, and the difference will be considerable for a t-distribution with a small number of degrees of freedom. So if the individual values have approximately such a distribution rather than a normal distribution, the values of α, power, and sample size will be off considerably from what they should be under the assumption of normality.

Mahnken (2009) discussed the determination of power and sample size when the distribution is unknown and provided a quasi-likelihood approach for use

when the variance as a function of the mean is known, or at least can be assumed. Of course, such sample size determination cannot be done exactly and the proposed methodology is based on asymptotic properties. Consequently, the estimates of sample size and power could be poor for small study designs, except, as the author points out, when the underlying distribution is normal.

3.5 ONE MEAN, EXPONENTIAL DISTRIBUTION

PASS has a routine for testing the mean of an exponential distribution and is the only major sample size determination software package that has this feature. It is perhaps also the only software or applet of any type with this capability. The exponential distribution has only one parameter, θ which is the mean of the distribution, so the null hypothesis would be a stated value of θ and the objective would be to determine the sample size for detecting a different value of θ that is considered important to detect.

To illustrate, let $\theta_0 = 2$ and $\theta_1 = 3$, with desired power of .90 and $\alpha = .05$ for a one-sided test. PASS gives various test criteria, including specifying a threshold value for $\hat{\theta}$, which is just the sample mean. The algorithm uses the theoretical result that $2n\bar{X}/\theta_0 \sim \chi^2_{2n}$(Epstein, 1960), with "$\sim$" read "is distributed as," followed by the statistical distribution. It follows that $\bar{X} \sim (\theta_0/2n)\ \chi^2_{2n}$. This result will be used in allowing us to see the threshold value. The software gives $n = 51$. With $2n = 102$, $\chi^2_{102,\,.05} = 126.74$, and $\theta_0/2n = 2/102 = 1/51$, it follows that $(\theta_0/2n)\ \chi^2_{2n,\,.05} = 126.74/51 = 2.48$, and the software gives 2.5. Thus, $\theta_0 = 2$ is rejected when $\theta_1 = 3$ using the reject criterion of $\hat{\theta} > 2.5$, with sampling without replacement and a fixed duration time of at least 12 time units, such that the study is terminated when that time has been reached. (By comparison, if the study duration time had been one unit, the necessary sample size would have been $n = 156$.)

3.6 TWO MEANS, KNOWN STANDARD DEVIATIONS—INDEPENDENT SAMPLES

As is the case for other tests covered in this chapter, there is an incongruity with some software in testing for the equality of two means with independent samples, as standard deviations of the two populations must be entered into software, but if the standard deviations were known, then the appropriate test would be the two-sample z-test, not the independent-sample t-test that some software assumes.

There are also some other problems with certain software, as when an independent-sample t-test is selected from the menu for the Power and Precision software but when the "variance is known" option is selected, the software indicates the selection with the message "z-test for two independent samples

with common variance" at the top of the screen. There is no such test as it isn't necessary for the two variances to be equal when the z-test is used.

Initially, it is not possible to enter two different values for σ_1 and σ_2 with that software, although it is possible to override it.

The form of the test statistic when Z is used is

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

with $\mu_1 - \mu_2$ usually equal to zero under the null hypothesis. For a one-sided test, assuming the use of $n_1 = n_2 = n$, the expression for n derived in the chapter Appendix, is

$$n = \frac{\left(\sigma_1^2 + \sigma_2^2\right)\left(Z_\alpha + Z_\beta\right)^2}{(\Delta - \Delta_0)^2} \tag{3.6}$$

with Δ denoting the difference, $\mu_1 - \mu_2$, that one wishes to detect, and Δ_0 is the difference under the null hypothesis, with Δ_0 usually but not necessarily being zero. For a two-sided test, Z_α is replaced by $Z_{\alpha/2}$. Although this simple formula modification will generally work, it won't necessarily work for the same reason that it won't work in the one-sample case, as discussed in Section 3.1.

For example, assume that $\sigma_1 = \sigma_2 = 1$, a two-sided test of H_0: $\mu_1 = \mu_2$ is to be performed with $\alpha = .05$, power $= .80$, and $\Delta = 1$. The computed value of n is 15.7, so $n = 16$ would be used for each sample size. This is the same solution that is obtained using PASS. If a one-sided test were used, with the same value of α and everything else also the same, then $n = 13$ for each group, which is what PASS gives. (Note that no solution is produced if the PASS user inadvertently selects the wrong side for the one-sided test, which would correspond to $\Delta = -1$ in this example.)

In this "variances known" case, we would hope that σ_1^2 and σ_2^2 are each estimated from a very large amount of data, which would justify the use of Z instead of t.

Simplified formulas are often given to industrial personnel to minimize the statistical expertise that is needed by the user. Kohavi and Longbotham (2001) gave the sample size formula

$$n = \frac{16\sigma^2}{\Delta^2} \tag{3.7}$$

for a t-test against a control, with σ^2 denoting the variance of the "overall evaluation criterion," and 16 is the number chosen to give power of .80, with $\alpha = .05$.

There is no indication how the 16 is obtained, however, and also no mention of a significance level. The authors indicated that replacing 16 by 21 will increase the power to .90. Again, no explanation, but the article reflects an obvious attempt to avoid technical details and give a simplified presentation. We can verify the 16 and 21 numbers if we let $\sigma_1^2 = \sigma_2^2 = \sigma^2$ in Eq. (3.6) and replace Z_α by $Z_{\alpha/2}$ for a two-sided test. Then the constants are 15.68, which would round to 16, and 21.015, which would round to 21. The latter is calculated as $(1^2 + 1^2)(1.95996 + 1.28155)^2 = 21.0148$. (The calculation of 15.68 is shown later.) This formula was indicated by Kohavi, Longbotham, Sommerfield, and Henne (2009) as being from van Belle (2008), who gave it as a rule-of-thumb that can be found in Chapter 2.

Kohavi et al. (2009) also mentioned the "more conservative" portable power formula of Wheeler (1974), which is $n = (4\, r\sigma/\Delta)^2$, with r denoting, in general, the number of levels of a factor in experimental design, so $r = 2$ here because there are two populations involved. Wheeler (1974) assumed the use of $\alpha = .05$ and power $= .90$, presumably for a two-sided test, although Wheeler did not indicate whether it is for a one-sided test or a two-sided test. It is obvious that Wheeler intended the power to be used in experimental design since there is only a single standard deviation symbol in the formula, whereas there are two such terms in Eq. (3.6). The following quote from Wheeler (1974, p. 193) is rather illuminating.

> We do not, however, think it always desirable to use the precise formulas because 1) their precision is in part illusionary due to poorly known parameter values, and 2) their richness produces client-consulting dialogues which are counter-productive: e.g., should the power be 0.90, 0.95, or 0.99?

Undoubtedly many statisticians and others would disagree with that position. A counterargument would be that it is preferable to start with the correct sample size expression and then show what the sample size would be for different parameter values. (This is illustrated in Section 3.11.) Furthermore, consultant–client relationships are best when the client has a reasonable knowledge of statistics, or at least a good aptitude for it. (I speak from the position of someone with many years of consulting experience.)

Wheeler's sample size formula was used by Kohavi, Henne, and Sommerfield (2007), with the authors pointing out that the sample size could be overestimated by 25% for large n. (Of course, an approximation error of that magnitude is generally unacceptable.) The formula was also criticized by Bowman and Kastenbaum (1974), which motivated a rejoinder by Wheeler (1975).

The approximation formula given by Lehr (1992) is essentially the same as the one later proposed by Kohavi and Longbotham (2001). The former is $n = 16\, s^2/d^2$, with s designating an estimate of the standard deviation, and d corresponding to Δ in Eqs. (3.6) and (3.7). As in the explanation of the Kohavi and Longbotham

(2001) formula, the "16" results from the fact that, if we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and assume use of a two-sided test with $\alpha = .05$ and power $= .80$, $2(1.96 + 0.84)^2 = 15.68$, with $Z_{\alpha/2} = 1.96$ and $Z_{.20} = 0.84$. So 15.68 is rounded to 16.

Thus, the approximation formula of Lehr (1992) can essentially be justified, but the Wheeler (1974) formula cannot be justified if it is used for a test of the equality of two means, as the constant is then 64 [obtained from $(4 * 2)^2$], which is much too large.

It is worth noting that although the use of these approximation formulas does simplify matters for scientists, they won't necessarily be satisfied with power of .80, nor should they be. So simplification does have a price.

3.6.1 Unequal Sample Sizes

Although equal sample sizes would be used in most applications, in some applications it might not be possible or practical to do so. Consider the example given in Section 3.6 but this time we won't assume that $n_1 = n_2 = n$. If two distinct sample sizes are to be used, a relationship must be specified since they can't be solved for individually, which wouldn't make any sense. We will enter $n_2/n_1 = 1.4$, and see how that affects the results. When PASS is used, the sample sizes obtained are 14 and 20 and the power is given as .8185. To obtain this solution in Power and Precision it is necessary to "link" the sample sizes. Accessing the option "N-cases" and then entering 10 and 14, respectively (so that the ratio is 1.4), in "Enter cases in ratio," results in the same solution as given by PASS. (Neither nQuery nor Lenth's applet has a procedure for testing two means with known standard deviations.)

It is worth noting the discussion in Campbell, Julious, and Altman (1995), who pointed out that if the sample size is computed under the assumption that $n_1 = n_2$ will be used but an unequal allocation is actually used, there will be a loss of power. They stated that the loss in power is only about 5% if the allocation ratio is actually 2:1, but drops off considerably beyond that point, with the loss being about 25% if the ratio is 5:1.

3.7 TWO MEANS, UNKNOWN BUT EQUAL STANDARD DEVIATIONS—INDEPENDENT SAMPLES

Technically, this case actually cannot be handled because power has to be computed with all necessary parameter values specified, including σ_1 and σ_2. If either they are not known or values are not assumed for them, then power cannot be computed. (Of course, parameter values are generally unknown, whether we assume known values for the purpose of determining sample sizes or not.)

When software is used, however, such as Power and Precision, one of the options is "*t*-test for independent samples with common variance." A standard deviation must be entered for one of the populations, with the standard deviations linked so that the same standard deviation is used for the other population, as required for the test. The sample sizes are also linked, so that the use of equal sample sizes is assumed. There is an obvious problem with doing this because if the standard deviations were known, a *z*-test would be used rather than a *t*-test, but sample size cannot be computed without specifying the standard deviations. Of course, the same thing happens with other software but PASS does allow the user to specify a range of values for the standard deviations with the output showing what the sample size would be for each value of $\sigma = \sigma_1 = \sigma_2$. (This would have to be done manually in Power and Precision, after the user has specified the increment size for σ. nQuery does not have the capability for providing results for multiple values of σ without the user having to enter each value individually and then running the program.)

Using multiple values of σ is somewhat more realistic than specifying a single value of $\sigma = \sigma_1 = \sigma_2$ and using the sample size that the software gives since the software is determining the sample size for the *t*-test that is based on the assumption that the common standard deviation is unknown. This option is available in SAS Software for this test. As shown in the documentation at `http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_power_a0000000970.htm`, there is the wording "you conjecture two scenarios for the common group standard deviation." This is the type of wording that is preferable for all sample size determination software that determines sample size under the assumption that a *t*-test will be used.

To see the effect of assuming the use of a *t*-test rather than the *z*-test that was assumed for the example that was used at the beginning of Section 3.6, let's return to that example. The objective was to detect a difference of one in the population means, with $\alpha = .05$, power $= .80$, and $\sigma_1 = \sigma_2 = 1$ for a two-sided test. The common sample size that was obtained was 16. For the *t*-test, the sample size that is obtained is 17, so there is a very slight increase in the sample size.

Of course, the assumptions that underlie the *t*-test should be met approximately before the test is used. This is another motivation for the use of an internal pilot study because the data from such a study could be used to test those assumptions. Should the assumptions appear to not be approximately met, the experimenter could then pursue a new direction, such as determining the sample size for the corresponding nonparametric test. Rasch, Kubinger, and Moder (2011) contended, however, that the assumptions should *not* be tested as doing so on the same data that will be used for the test alters the Type I and Type II errors in an unknown way. Instead, they recommended routinely using the Welch (nonparametric) test.

That is not a strong reason to avoid testing the assumptions, however, especially since the amount of contamination is unknown but would likely be slight, especially if an internal pilot study was less than half of the total study.

3.7.1 Unequal Sample Sizes

Although equal sample sizes are generally desirable, that isn't always possible or practical, as illustrated by the following example.

■ EXAMPLE 3.4

Newcombe (2001) referred to a study by Heasman et al. (1998) in which they found that the mean toothbrushing force at baseline was 220 grams for boys and 181 grams for girls. Motivated somewhat by this study, Newcombe (2001) stated: "Suppose we decided that in a new study, we want an 80 percent power to detect a difference of 30 grams as statistically significant at the 5 percent level. Based on the existing study, it seems reasonable to assume a SD of 130g. Suppose that, as in the published study, we expect to recruit twice as many girls as boys."

The anticipated difference in sample sizes can easily be handled by most sample size determination software. Using PASS and assuming a common standard deviation of 130, $\alpha = .05$ for a two-sided test, desired power = .80, an expected difference of 30 grams, and a 2:1 ratio for sample sizes, PASS gives 223 and 446 for the two sample sizes with power = .8017, whereas both nQuery and Power and Precision give 222 and 444 for the sample sizes. When these sample sizes are entered in nQuery, the latter displays power = .8004. The power value given by Power and Precision is almost identical, as the output shows power = .8005. (MINITAB does not have a two-sample *Z*-test and does not permit unequal sample sizes in solving for sample size with its two-sample *t*-test.)

Newcombe (2001) used a rather crude approach to arrive at 221 and 442 for the two sample sizes, which used an initial guess for the sample sizes (obtaining 100 and 200), computing the value of the *t*-statistic if those sample sizes had been used (1.88), then multiplying the 100 and 200 by $(2.80/1.88)^2$ Although not explained in the article, the $2.80 = 1.96 + 0.84 = Z_{\alpha/2} + Z_{\beta}$—a large-sample approximation since the use of a *t*-test is being assumed, not a *Z*-test. Actually, Newcombe (2001) should have given the smaller sample size as 222 instead of 221 since the computed value is 221.077 and the sample size is always rounded up to the next integer. The 223 given by PASS is due to the search procedure employed by PASS, which ensures that the power is at least equal to the desired power. With 222 and 444 as the sample sizes, PASS gives the power as .79994, which of course is essentially .80, although technically it is less than .80. The power jumps to .8017 for sample sizes of 223 and 446. Thus, we can think of either pair of sample sizes as being "correct."

The bottom line is that the two-stage approach of Newcombe (2001) does work, although it is unnecessary now and also then since software can be used to obtain the sample sizes with less effort. Of course, we might question why a two-sided test was being assumed since the example was motivated by a study that showed results in a particular direction.

Other software that can be used when it is desired to use unequal sample sizes includes PASS, which gives the same solution as nQuery. The latter does not give the actual power in its sample size determination output, but the power can be obtained after the sample sizes are determined by entering those sample sizes in addition to the other required input. Doing so results in nQuery giving the power as .8195. This differs slightly from the power given by PASS, which is .8185. Power and Precision does not permit the direct determination of the sample sizes unless n_2/n_1 is a specified ratio. The user can manipulate the two sample sizes individually, however, and see what effect that has on power. Specifying $n_1 = 15$ and $n_2 = 21$ gives .8195 for the power, in agreement with nQuery. Lenth's applet gives the same value for power for this combination of sample sizes. ■

3.8 TWO MEANS, UNEQUAL VARIANCES AND SAMPLE SIZES—INDEPENDENT SAMPLES

Schouten (1999) proposed a method for determining sample sizes in testing the equality of two means for unequal variances and unequal sample sizes that the experimenter wishes to use. Such a method will often be needed because unequal sample sizes exacerbate the effect of unequal variances. The sample size formula is

$$n_2 = (Z_{\alpha/2} + Z_\beta)^2 \frac{(\tau + \gamma)\,\sigma_2^2}{\gamma\,(\mu_2 - \mu_1)} + \frac{(\tau^2 + \gamma^3) Z_{\alpha/2}^2}{2\gamma\,(\tau + \tau)^2}$$

$$n_1 = \gamma n_2$$

$$n = n_2 + n_1$$

with γ denoting the ratio of the sample size from population 1 to the sample size from population 2, τ is the ratio of the population 1 variance to the population 2 variance, and σ_2^2 is the population 2 variance. Note that Z is used, which means that approximate normality is assumed. The need for a method such as given by Schouten (1999) is not as great when the sample sizes are equal, although Schouten's method might still be used.

3.9 TWO MEANS, UNKNOWN AND UNEQUAL STANDARD DEVIATIONS—INDEPENDENT SAMPLES

The obvious question is: "If we assume that the standard deviations are unknown, how do we know that they are unequal?" The answer to this apparent conundrum is that the assumption of unknown standard deviations (which technically is

always the case, regardless of what we assume) results in the *t*-test being used by software, but the hypothesis of equality of the variances should be tested, as illustrated in Section 3.11.

3.10 TWO MEANS, KNOWN AND UNKNOWN STANDARD DEVIATIONS—DEPENDENT SAMPLES

Even though computationally the sample from each of two populations is collapsed into a single set of differences between each pair when the data are paired, the sample size to be used and the power associated with the selected sample size should logically depend on the variability of the data in each of the populations and the correlation between the random variables being used for each of the two populations.

This can be seen by examining the denominator of the test statistic. For simplicity, we will initially assume that the two populations each have a normal distribution and that the population standard deviations are known. We will let d_i represent the ith difference in a set of n paired differences. The test statistic is then

$$Z = \frac{\bar{d}}{\sigma_d}$$

If we let the observations from the first population be represented by y_1 and the observations from the second population be represented by y_2, then $d = y_1 - y_2$ and $\sigma_d = \sigma_{y_1 - y_2}$. Since

$$\sigma_{y_1 - y_2} = \sqrt{\sigma_{y_1}^2 + \sigma_{y_2}^2 - 2\sigma_{y_1 y_2}} = \sqrt{\sigma_{y_1}^2 + \sigma_{y_2}^2 - 2\rho_{y_1 y_2}\sigma_{y_1}\sigma_{y_2}}$$

with $\sigma_{y_1 y_2}$ and $\rho_{y_1 y_2}$ denoting the covariance and correlation, respectively, between y_1 and y_2, it is obvious that the variability of each population and the correlation between the two random variables will influence the power of the test since they determine the denominator of the test statistic.

We can see this more formally if we derive the sample size expression, proceeding analogous to what was done for a one-sample mean. That is, we will assume that the null hypothesis is $\mu_d = 0$ and that the alternative hypothesis is $\mu_d = \mu^* > 0$. Then, still assuming known values for the population variances, covariance, and correlation, the development of the expression for n would have been exactly the same as the development in Example 2.1, resulting in Eq. (2.3), so that will not be repeated here. The end result is

$$n = \left[\frac{(Z_\alpha + Z_\beta)\sigma_d}{\mu_d}\right]^2$$

■ EXAMPLE 3.5

To illustrate, let's assume that the data are paired because the same people are going to be used for "before" and "after" readings, with "after" perhaps being readings after some type of treatment program. With $d = y_1 - y_2$ and the null hypothesis being that $\mu_d = 0$, assume that we wish to detect $\mu_d = 1$. We will let $\alpha = .05$, assume that the alternative hypothesis is $\mu_d > 0$, and use the customary .80 as the selected power value. We will also assume $\sigma_{y_1} = \sigma_{y_2} = 5$ and $\rho_{y_1 y_2} = .80$. Then $\sigma_d = \sqrt{25 + 25 - 2(.80)(25)} = \sqrt{10} = 3.162$. Thus,

$$n = \left[\frac{(1.645 + 0.84)\sqrt{10}}{1}\right]^2 = 61.75 \text{ (round to 62)}$$

■

When PASS is used, the user must input the 3.162 for the standard deviation and check "known standard deviation." When this is done, the output gives $n = 62$, in agreement with the above calculation. When the standard deviation is not assumed to be known, PASS gives the sample size as 64. (MINITAB does not give an option regarding the standard deviation being known or not and simply gives the sample size as 64.) The larger sample size is due to the fact that t would be used rather than z. The sample size is thus computed as

$$\begin{aligned} n &= \left[\frac{(t_{\alpha, n-1} + t_{\beta, n-1})\hat{\sigma}_d}{\mu_d}\right]^2 \\ &= \left[\frac{(t_{\alpha,64-1} + t_{\beta,64-1})\hat{\sigma}_d}{\mu_d}\right]^2 \\ &= \left[\frac{(1.6694 + 0.847364)\,3.162}{1}\right]^2 \\ &= 63.33 \end{aligned}$$

so $n = 64$ would be used. (Of course, here the known solution for n is being used to obtain that solution; this is simply for illustration.)

In this example, the two standard deviations and the correlation were assumed to be known. Generally, they won't be known and will have to be estimated in some manner. For before and after measurements, especially the latter, and for other types of paired data, it may not be possible to estimate standard deviations. An experimenter may have to assume that the "after" variability will be the same as the "before" variability, with any change being only a change in the mean. Depending on the nature of the experiment, this could be a very shaky assumption. Furthermore, how does one estimate the correlation between "before" and "after"

when "after" hasn't even occurred? The assumption of only a mean change (at most) and no change in variability implies a correlation of 1.0 between the before and after measurements, which could also be a very shaky assumption. (If the variability does change, then that will be reflected in the variability of the differences when they are formed and the value of the test statistic is computed.)

Thus, some very shaky assumptions could be made when sample size is determined for a paired-t test, which the test would have in common with other statistical methods when sample size is determined.

When variances are assumed to be unknown, software such as Power and Precision, MINITAB (Release 16), and Lenth's applet tacitly assume that the standard deviations and correlation *have been estimated* from the same number of pairs as the algorithm computes, as the sample size is computed as if a paired-t test was being used with the degrees of freedom for the t-statistic being that which results from the computed sample size. For example, for the previous example, the Power and Precision software gave $n = 64$ as the solution for n, as does MINITAB. Similarly, when Lenth's applet is used (for which the user must compute an estimate of σ_d and enter that), the solution is $n = 64$ and the power is given as .8038. It is clear from the menu that the sample size is being computed for a t-test. If we do the hand computation using the "known" value of n to obtain the t-values, we would obtain $n = (1.6694 + 0.847364)^2(10) = 63.34$, so $n = 64$ would be used.

Of course, it is unrealistic to assume that the variances have already been estimated from two paired samples each of size $n = 64$. It would be even more unrealistic to assume that after the experiment has been performed, s_d will be equal to the value that we have assumed for σ_d.

In this example, the difference between $n = 62$ and $n = 64$ is apt to be inconsequential, but strictly speaking, the sample size should be computed based on how the standard deviations and correlation are estimated. If we assume that the difference, d, has a normal distribution and the standard deviations and correlation are estimated (perhaps poorly) without the use of data, the appropriate statistic is Z, not t.

Unfortunately, a "z-test" for paired data, as available in MINITAB and Lenth's applet, doesn't seem to be an option in some software. Although it is not part of the menu and thus might be easily overlooked, the Power and Precision software does have an option for each of its t-tests to convert the test to a z-test with the assumption of known variance(s), just as it does in the one-sample case. For the current example, if the option of a known σ_d is selected and 3.162 is entered for the assumed known value, the software gives $n = 62$, in agreement with the value obtained by hand computation for this example.

For all practical purposes, this distinction between "known" and "unknown" population standard deviations is a superficial one at best because population parameters are never really known. When the standard deviations are assumed unknown, values for them must still be entered for the algorithm to be able to

compute sample size. When values must be entered into an algorithm for the purpose of determining the necessary sample size, the parameter values are being guessed, both for the "known" case and the "unknown" case. If such guesses are being made based on, say, a very large amount of data accumulated over time, then a *t*-test would be inappropriate.

In teaching an online sample size determination and power course a few years ago, I encountered a student who was interested in not just a confidence interval on the mean difference for treatment minus control, but also wanted separate confidence intervals for treatment and control. Such analyses are generally not done but under certain conditions separate confidence intervals could provide important information. For example, although no mean value for "control" is assumed when a paired-*t* test is performed, such a value is assumed in determining the sample size to use. Perhaps that assumed value is incorrect/unrealistic. A confidence interval constructed for the mean might just barely contain the assumed mean, which would cast some doubt on the latter. If there is some evidence that the control mean is greater than the assumed control mean, the paired-*t* test might not give a significant result even though a confidence interval for the treatment mean may suggest that the treatment is very beneficial.

A power analysis for a paired-*t* test using SAS can be performed using PROC Power, with the code given and the analysis illustrated at `http://www.ats.ucla.edu/stat/sas/dae/t_test_power3.htm`.

3.11 BAYESIAN METHODS FOR COMPARING MEANS

There are various articles in the literature on Bayesian methods of sample size determination for testing hypotheses about means. This is a natural consideration because some information regarding population parameters must be used as input when sample sizes are determined using software. Furthermore, it is almost essential to first use a pilot study to obtain some data for parameter estimation, so some type of prior information is essential.

Wang, Chow, and Chen (2005) proposed a very simple Bayesian approach for sample size determination in comparing two means. Since their method was proposed for clinical research, it is discussed in Section 7.8. Joseph and Bèlisle (1997) provided Bayesian sample size determination methods for population means and differences between means for normal distributions.

3.12 ONE VARIANCE OR STANDARD DEVIATION

There is often interest in testing a variance or standard deviation, hoping to see if there has been a reduction since reduced variability is desirable for virtually anything that is measured, especially a process characteristic in manufacturing.

There might seem to be a major problem in trying to test a hypothesis on σ^2 or construct a confidence interval for it, however, because when normality is assumed, $\text{Var}(s^2)$ is a function of σ^4. Thus, a practitioner would seem to need a good estimate of σ or σ^2 in order to determine the sample size.

That isn't the case, however, as can be seen with the following example. Assume that we want to test H_0: $\sigma^2 = 2$ vs. H_0: $\sigma^2 < 2$, using $\alpha = .05$. We want the power to be .90 if $\sigma^2 = 1$. Using PASS, if we enter these numbers, the output has $n = 39$ as the sample size, which gives power = .90423. We can verify this as follows. $(n-1)\, s^2/\sigma^2$ has a chi-square distribution with $(n-1)$ degrees of freedom. Let $X = (n-1)\, s^2/2$. Then $X \sim \chi^2_{n-1}$. We may show using statistical software $\chi^2_{38,\,.05} = 24.8839$. Reducing σ^2 to 1 has the effect of doubling the value of X. Since $2X \sim 2\chi^2_{n-1}$, we thus want $P(X < 2\chi^2_{38,\,.05}) = .90$. It can be shown that $P(X < 2x^2{}_{38,\,.05} = 2(24.8839) = 49.7678) = .90423$, which is the value that PASS gives for the power.

So the solution can easily be seen and verified as correct, but a straightforward derivation is not possible because the distribution of the test statistic depends on the sample size, which is what we are trying to determine! Consequently, an exact expression for the sample size cannot be obtained and an iterative solution is necessary.

Approximations have been proposed, which might appeal to practitioners who don't have access to sample size determination software and don't need an exact solution for the sample size. Bonett and Al-Sunduqchi (1994) examined the performance of an approximation given by Duncan (1986, p. 401) and found that it worked well only when $\alpha = \beta$. Since values of α greater than .05 are seldom used, and similarly power values of at least .95 are typically not used, this requirement generally will not be met. They gave the following approximation:

$$n = \left(\frac{Z_{\alpha/2} + Z_{\beta}}{Z_{p1}} \right)^2$$

with $Z_{p1} = (|\ln(\sigma^2) - \ln(k)|)/2^{1/2}$ and $Z_{\alpha/2}$ replaced by Z_{α} for a one-sided test. Here k denotes the hypothesized value of σ^2, with σ^2 in the expression for Z_{p_1} denoting the actual variance, and $Z_{\alpha/2}$ and Z_{β} are as defined in Section 2.1. Applying this approximation to the current example (and using Z_{α} in place of $Z_{\alpha/2}$ since it is a one-sided test), we obtain

$$\begin{aligned} n &= \left(\frac{1.645 + 1.28}{0.4901} \right)^2 \\ &= 35.6 \end{aligned}$$

so $n = 36$ would be used, differing from the solution of $n = 39$ obtained using PASS. Bonett and Al-Sunduqchi (1994) recommended the sample size obtained

with this approximation be averaged with the sample size obtained using the Duncan approximation. The latter is

$$n = \left(\frac{k^{1/2} Z_{\alpha/2} + \sigma Z_{\beta}}{Z_{p2}} \right)^2$$

with $Z_{p2} = 2^{1/2} | k^{1/2} - \sigma |$. After again replacing $Z_{\alpha/2}$ by Z_{α}, we obtain a sample size of 37.9, so 38 would be used.

As stated, Bonett and Al-Sunduqchi (1994) recommended averaging the sample size obtained using this approximation with the value obtained using Duncan's approximation. If we average 35.6 and 37.9, we obtain 36.75, so 37 would be used, slightly less than the 39 obtained using PASS. So the recommended approximation is a bit lacking for this example but certainly adequate if only a ballpark figure is needed.

If the objective is to simply estimate a standard deviation as a percentage of its true value, Greenwood and Sandomire (1950) addressed the determination of sample size for this purpose.

3.13 TWO VARIANCES

It is often necessary to test the equality of two variances, as procedures such as the independent-sample *t*-test assume that the two variances are equal. Therefore, equality of the two variances should be tested before the *t*-test is used. There are also many scenarios when a *t*-test will not be used for which it is desirable to test equality of variances, such as when a supposedly improved manufacturing process has been installed and there is a desire to see if this has resulted in more uniform manufactured products being produced (i.e., smaller variance).

Since two variances will hardly ever be equal, there is the obvious question of what departure from equality should the experimenter attempt to detect. If equal sample sizes will be used and normality can be assumed, there can be a moderate difference in two variances without causing any problems with the *t*-test. Markowski and Markowski (1990) stated: "Specifically, for equal sample sizes, the *t* test is generally robust and hence no preliminary test is needed. Boneau (1960) reported exceptions to this robustness, such as with one-tailed alternatives when sampling from skewed distributions or with very small significance levels." Thus, if the use of equal sample sizes is planned, most of the time it won't be necessary to test the equality of variances assumption, but it would still be a good idea to try to gain insight into the shape of the distributions for the two populations, as severe nonnormality could create a problem.

Let's assume equal sample sizes and slight-to-moderate nonnormality for the following example. Recognizing that the appropriate sample size will be

determined by the ratio of the two variances, not their individual values, let's consider ratios of 3:1, 5:1, 7:1, and 9:1, as in 10 for the first population variance and 30, 50, 70, and 90 for the second population variance. Using $\alpha = .05$ for a two-sided test and desired power of .90, PASS gives the sample sizes from each population as 37, 19, 14, and 11, respectively. This assumes use of an *F*-test, however, which is very sensitive to departures from normality, as was first noted by George Box, who also noted the *t*-test is robust to nonnormality. Unfortunately, software developers have generally not provided an option for determining sample size using something other than the *F*-test, with no option provided by PASS or G*Power, and no test of any kind is available in Power and Precision or nQuery. MINITAB does, however, have Levene's test as an option. The corresponding sample sizes produced by MINITAB when the Levene test option is selected are 46, 25, 19, and 16, respectively, which can be seen to differ considerably from the sample sizes for the *F*-test under the assumption of normality. It is generally better to use a test of the equality of two variances that does not assume normality. So here it would be wise to be guided by the sample sizes needed for Levene's test. [See Gibbons (2000) for the determination of sample sizes for Levene's test.] Markowski and Markowski (1990) discouraged the use of an *F*-test even under normality, stating: "However, our study has shown that even when sampling from a normal distribution, the *F* test is unlikely to detect many situations where the *t* test should be avoided. In conclusion, we feel the *F* test is flawed as a preliminary test."

Although a test for equality of variances is typically performed when the two samples are independent, Pitman (1939) proposed a test for equality of variances for paired, normally distributed data that was based on the correlation between the sums and differences within the pairs. Bogle and Hsu (2002) proposed three methods for testing two population variances with paired data and illustrated the method with the highest power in an example involving bilirubin tests.

3.14 MORE THAN TWO VARIANCES

There is often a need to test for the equality of more than two variances, such as when Analysis of Variance (ANOVA) is used with unequal sample sizes. Wludyka, Nelson, and Silva (2001) provided power curves for the Analysis of Means (ANOM) for variances, which, as they stated, can be helpful in determining sample size when ANOM for Variances is used to test for homogeneity of variances. Table B.5 and Appendix A in Nelson, Wludyka, and Copeland (2005) can also be used for determining sample size and corresponding power.

3.15 CONFIDENCE INTERVALS

Determining sample size for a confidence interval for a single mean was discussed in detail and illustrated in Section 2.4, and also illustrated in Example 3.2. Since

most sample size determination software is focused primarily on hypothesis tests, there is slightly less software to choose from for determining sample size for confidence intervals, as indicated previously. MINITAB does have that capability, however, as does PASS.

To illustrate, in Section 2.4 the sample size was determined for a 95% confidence interval for μ, assuming $\sigma = 2$, and E (the maximum error of estimation) $= 1$. Both hand computation and the use of PASS resulted in $n = 16$. When MINITAB is used and the same information is inputted, $n = 16$ is the output. Since σ is unknown, it is a good idea to input multiple values of σ and see what n is for each value. So for this example 1.5, 2.0, and 2.5 might be used and when this is done in PASS, the sample sizes are 9, 16, and 25, respectively. Of course, as discussed in Section 2.2, there are ways to estimate σ without data and PASS and nQuery, in particular, can be useful in this regard.

This is not possible in MINITAB, however, except for individually solving for n for each desired value of σ, but it is possible to input multiple values of E at one time. PASS does allow both to be inputted at one time, however, so if three values are inputted for σ and three values inputted for E, the output will show nine sample sizes—one for each combination.

3.15.1 Adaptive Confidence Intervals

The term "adaptive confidence intervals" has been used in the literature, such as in Hartung and Knapp (2010). This refers to a multistage approach, such as when an internal pilot study is used. Hartung and Knapp (2010) pointed out that this is an old problem, with Stein (1945) having provided a two-stage procedure, with the sample size for the second stage based on the results from the first stage. Seelbinder (1953) showed how to select the sample size for the second stage when there is some prior information on σ^2.

3.15.2 One Mean, Standard Deviation Unknown—With Tolerance Probability

When σ is unknown and approximate normality is assumed, the t-distribution is used. Of course, a value for σ must be entered and, as in every sample size determination situation, the inputted value will almost certainly not be the same as the true value. Both PASS and nQuery have two routines for determining sample size for μ with σ unknown—one that accounts for the variability in the estimate of σ and one that does not. The former is called a "coverage correction" by nQuery and a "tolerance probability" by PASS, with nQuery recommending use of a coverage correction (probability) of .80. PASS does not give a specific recommendation but the default value is .90. When this option is not used, the sample size that is determined is valid only if σ is not greater than the assumed value. If it is greater, then the sample size that has been computed will be too small.

If σ is being estimated from a previous sample, the adjustment, as given originally by Harris, Horvitz, and Mood (1948), produces

$$n = \left(\frac{t_{\alpha/2,n-1}\hat{\sigma}}{E} \right)^2 F_{1-\gamma,n-1,m-1} \tag{3.8}$$

with $1-\gamma$ denoting the probability that the distance from the mean to the $100(1-\alpha)\%$ two-sided confidence limits, E, will be less than the specified value, and m is the size of the previous sample used in estimating σ.

When no previous sample is used for estimation, the sample size expression, as given by Kupper and Hafner (1989), for example, is

$$n = \left(\frac{t_{\alpha/2,\,n-1}\sigma^*}{E} \right)^2 \left(\frac{\chi^2_{1-\gamma,\,n-1}}{n-1} \right) \tag{3.9}$$

with $1-\gamma$ and E having the same representation as in the previous formula and $\hat{\sigma}$ assumed to be the population value. The first fraction in Eq. (3.9) results from modifying the sample size expression given in Eq. (2.7) by replacing $Z_{\alpha/2}$ by $t_{\alpha/2,\,n-1}$ (since σ is unknown) and replacing σ by s^*, the upper $100(1-\gamma)\%$ prediction bound for s, the sample standard deviation for the future sample whose sample size is being determined. This gives

$$n = \left(\frac{t_{\alpha/2,\,n-1}s^*}{E} \right)^2 \tag{3.10}$$

The second fraction is obtained from the upper limit of a $100(1-\gamma)\%$ prediction bound for s, which is

$$s^* = \sigma^* \left(\frac{\chi^2_{1-\gamma,\,n-1}}{n-1} \right)^{1/2}$$

with σ^* denoting the true value of σ. Substituting this expression for s^* in Eq. (3.10) then produces Eq. (3.9).

Of course neither Eq. (3.8) nor Eq. (3.9) is a closed-form expression since n is on the right side of each formula. Thus, each one must be solved numerically. This would serve as an impediment to the use of coverage probability if this option were not available in software, but as indicated earlier in this section, it is available in software.

■ EXAMPLE 3.6

To illustrate, consider again a 95% confidence interval for μ with $E = 1$ and this time assume that, using a previous sample of size $m = 40$, $\hat{\sigma} = 2$. Using the default coverage probability in PASS of .90, the software gives $n = 27$. If we had (numerically) calculated the sample size using Eq. (3.8), we would have obtained

$$\begin{aligned} n &= \left(\frac{t_{\alpha/2,n-1}\hat{\sigma}}{E}\right)^2 F_{1-\gamma,n-1,m-1} \\ &= \left(\frac{(2.05553)(2)}{1}\right)^2 (1.5660) \\ &= 26.4667 \end{aligned}$$

so $n = 27$ would be used. (Note that here I am using the known result from PASS of $n = 27$ to show that the formula gives that result.) Note also that we round up to the next integer value just as we do when a hypothesis is being tested, but for a different reason. We want the difference to be at most equal to the target value, just as we want the power to be at least equal to the target value for the hypothesis test. We usually won't be able to hit the target value exactly, however, so we have to decide whether we prefer a sample size that is slightly under the target value or slightly over the target value, with the former generally preferred.

The sample size is relatively insensitive to the size of the previous sample, provided it is large, as $m = 1000$ gives $n = 24$.

When the distribution of σ is assumed to have a mean of 2 (note that this is not the same as assuming that σ is known, as in that case all of the probability would be at the known value), $n = 24$, which is the same solution obtained using nQuery. Using Eq. (3.9) and still assuming a 95% confidence interval with $1 - \gamma = .90$ and $E = 1$, we obtain

$$\begin{aligned} n &= \left(\frac{t_{\alpha/2,n-1}\hat{\sigma}}{E}\right)^2 \left(\frac{x^2_{1-\gamma,n-1}}{n-1}\right) \\ &= \left(\frac{(2.06866)(2)}{1}\right)^2 \left(\frac{32.0069}{23}\right) \\ &= 23.8207 \end{aligned}$$

so $n = 24$ would be used (again using the known result to illustrate the formula). (Note that although n appears in the denominator of the second fraction, there is no point in trying to write the sample size formula in a cleaner fashion since

n also appears in the numerator of that fraction and the numerator of the other fraction.) ■

3.15.3 Difference Between Two Independent Means, Standard Deviations Known and Unknown—With and Without Tolerance Probability

If we consider the simplest case of constructing a confidence interval for $\mu_1 - \mu_2$ with σ_1 and σ_2 assumed known and equal sample sizes, n, to be used, the sample size expression is obtained by using the expression for the Z-statistic given in Section 3.5. Specifically, the expression for the halfwidth of a $100(1-\alpha)\%$ confidence interval, with equal sample sizes, is

$$Z_{\alpha/2}\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}} = E$$

Solving this expression for n gives

$$n = \frac{Z_{\alpha/2}^2\left(\sigma_1^2 + \sigma_2^2\right)}{E^2}$$

For example, if $\sigma_1^2 = 9$, $\sigma_2^2 = 16$, $E = 2$, and a 95% confidence interval is to be constructed,

$$\begin{aligned} n &= \frac{(1.96)^2\,(9+16)}{2^2} \\ &= 24.01 \end{aligned}$$

Somewhat oddly, neither PASS nor nQuery has the capability to solve for the sample size in this the simplest of the two-sample mean cases.

If we assume that both standard deviations are unknown, but equal, we would then use a pooled-t test if we could assume approximate normality for the two populations. For the sake of comparison with the preceding example, we will use 3 and 4 as estimates of σ_1 and σ_2, respectively, and again use $E = 2$. We would expect the common sample size to be somewhat larger than 24 and, using PASS, the result is $n = 26$.

Thus,

$$E = t_{\alpha/2,2n-2}s_p\sqrt{\frac{2}{n}}$$

with s_p denoting the pooled standard deviation, which here is $\sqrt{(9+16)/2}$. Therefore,

$$\begin{aligned} n &= \frac{2t^2_{\alpha/2,2n-2}s^2_p}{E^2} \\ &= \frac{2\,(2.00856)^2\,(12.5)}{2^2} \\ &= 25.2145 \end{aligned}$$

so $n = 26$ would be used, as given by PASS. When nQuery is used, however, the solution given is $n = 25$. (MINITAB cannot be used for this since its capability for determining sample sizes for confidence intervals does not extend to two means.)

Both PASS and nQuery have the capability for a confidence interval for the difference of two means with a coverage probability. As in the one-sample case, there is a choice to be made between the common standard deviation being estimated from a prior pair of samples or not. Thus, the relevant equations are extensions of Eqs. (3.8) and (3.9) to the two-sample case.

First, if we use the same information as in the preceding example and specify a coverage probability of .90 with σ *not* estimated from a previous pair of samples, nQuery and PASS each give the sample size as $n = 32$. Of course, we expect the common sample size to be larger than in the case where the uncertainty in the estimates of the common standard deviation was not factored in, and we would expect the sample size to be larger than 32 if we assume that a prior pair of samples were used. Making that specification, and assuming 30 for each of the two prior sample sizes, PASS gives $n = 35$ but nQuery does not have this sample size capability, just as it does not in the one-sample case. (Of course, the sample size that PASS gives will approach 32 if we let the prior sample sizes be larger.)

For the case of standard deviations unknown with coverage probability, if a previous sample from each population is used to estimate each standard deviation and we assume both those samples and the samples to be taken from each population are equal, the sample size formula is

$$n = \frac{2t^2_{\alpha/2,2n-2}s^2_p F_{1-\gamma,2n-2,2m-2}}{E^2} \tag{3.11}$$

with $1 - \gamma$, s_P, and E as previously defined, and similarly for n and m. If a previous sample from each population is not available, then

$$n = \frac{t^2_{\alpha/2,2n-2}s^2_p\, x^2_{1-y,2n-2}}{E^2\,(n-1)} \tag{3.12}$$

To illustrate Eq. (3.11), PASS gives $n = 35$ as the sample size when a previous sample of size 30 has been obtained from each population. Making the appropriate substitutions in Eq. (3.11), we obtain

$$\begin{aligned} n &= \frac{2t^2_{\alpha/2,2n-2}s^2_p F_{1-\gamma,2n-2,2m-2}}{E^2} \\ &= \frac{2\,(1.99167)^2\,(12.5)\,(1.38113)}{4} \\ &= 34.24 \end{aligned}$$

so $n = 35$ would be used, as given by PASS.

As indicated, both PASS and nQuery give $n = 32$ when σ is *not* estimated from a sample taken from each population. To illustrate how the sample size would be computed using Eq. (3.12), we have

$$\begin{aligned} n &= \frac{t^2_{\alpha/2,2n-2}s^2_p x^2_{1-\gamma,2n-2}}{E^2\,(n-1)} \\ &= \frac{(1.99897)^2(12.5)(76.6302)}{4(31)} \\ &= 30.8675 \end{aligned}$$

so $n = 31$ would be used. Thus, the computed sample size is one unit less than the solution given by PASS and nQuery, but the PASS output shows that the actual value of E using $n = 32$ is 1.964 rather than the desired value of 2.0, compared to $E = 1.98517$ when $n = 31$ is used. That is, both PASS and nQuery give a solution with a smaller error of estimation than requested, which means a narrower confidence interval than requested and a narrower confidence interval than would result from the use of $n = 31$. For a given degree of confidence and a given coverage probability, the narrower the interval the better, but the slightly narrower interval that would result from using the PASS and nQuery results comes at the expense of a larger sample size.

3.15.4 Difference Between Two Paired Means

The hypothesis test for two paired means was discussed in Section 3.9. The form of a $100(1-\alpha)\%$ confidence interval for μ_d is, again assuming that σ_d is known, $\bar{d} \pm Z_{\alpha/2}\sigma_d/\sqrt{n}$, with n denoting the number of differences, d, and, equivalently, the common size of each paired observation in the paired samples.

Then D, the halfwidth of the interval, is $DZ_{\alpha/2}\sigma_d/\sqrt{n}$. Solving for n gives $n = Z^2_{\alpha/2}\sigma^2_d/D^2$. To illustrate, if $\sigma_d = 4$ and $D = 1$ and we want a 95% confidence for μ_d, we obtain $n = (1.96)^2(4)^2/1^2 = 61.4656$, so $n = 62$ would be used, which

is the solution given by PASS. If σ_d is not assumed to be known, but is estimated from a pilot study or some type of previous study or estimate, the formula is $n = t^2_{\alpha/2,n-1}\sigma_d^2/D^2$, which would have to be solved numerically, as is the case whenever the sample size expression is a function of a t-value. PASS gives $n = 64$ and using the formula with this known result we have $n = (1.99962)^2(16)/1 = 63.9757$, so $n = 64$ would be used.

If sample size is to be determined with a coverage probability of .90 with the same information as in the previous example, except that σ_d is estimated to be 4 from a previous sample of size $m = 30$, PASS gives $n = 96$. Since a paired-t test becomes effectively a one-sample t-test once the differences, d, have been formed, Eq. (3.8) applies when σ_d has been estimated from a prior sample, as in this case. Therefore,

$$\begin{aligned} n &= \left(\frac{t_{\alpha/2,n-1}\hat{\sigma}}{D}\right)^2 F_{1-\gamma,n-1,m-1} \\ &= [(1.98525)(4)]^2(1.51937) \\ &= 95.8107 \end{aligned}$$

so $n = 96$ would be used, in agreement with PASS. If a previous sample is not used and the user is willing to assume that future samples will have a standard deviation of 4 (a possibly unreasonable assumption), PASS gives $n = 77$ as the solution. Here Eq. (3.9) applies so that

$$\begin{aligned} n &= \left(\frac{t_{\alpha/2,n-1}\sigma^*}{D}\right)^2 \left(\frac{x^2_{1-\gamma,n-1}}{n-1}\right) \\ &= [(1.98525)(4)]^2 \left(\frac{92.1662}{76}\right) \\ &= 76.4731 \end{aligned}$$

so $n = 77$ would be used, in agreement with PASS.

3.15.5 One Variance

The expression for a two-sided confidence interval for σ^2 is obtained starting from

$$P\left(x^2_{n-1,\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq x^2_{n-1,1-\alpha/2}\right) = 1-\alpha$$

From this starting point we simply need to manipulate the form so that σ^2 will be in the middle of the double inequality. Doing the appropriate algebra produces

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1-\alpha \tag{3.13}$$

with the lower limit of the $100(1-\alpha)\%$ confidence interval being on the left side of the double inequality and the upper limit being on the right side of the double inequality.

This is an example of an asymmetric confidence interval because the distances from the sample variance, S^2, to each confidence limit are unequal. Therefore, it wouldn't make any sense to use a value for D, as with the preceding confidence intervals. Accordingly, with PASS the user specifies the width, W, of the confidence interval, which is the upper limit minus the lower limit.

Assume that we want the width of a 99% confidence interval to be $W = 8$ and we will assume that future samples will have $S^2 = 10$. PASS gives $n = 95$ as the solution. Taking the difference between the two limits in Eq. (3.13) and performing the necessary algebra, we obtain

$$\begin{aligned} n &= 1 + \frac{W\chi^2_{n-1,1-\alpha/2}\chi^2_{n-1,\alpha/2}}{S^2\left(\chi^2_{n-1,1-\alpha/2} - \chi^2_{n-1,\alpha/2}\right)} \\ &= 1 + \frac{8\,(133.059)\,(62.4370)}{10\,(133.059 - 62.4370)} \\ &= 95.11 \end{aligned}$$

Here we would not round up because with $n = 95$ the actual width of the interval is 7.991, as reported by PASS, whereas if we had used $n = 96$, the width would have been 7.943, which is further away from the target width than when $n = 95$ is used. We round up for hypothesis testing so that the power is at least equal to the desired power, but of course there is no power involved in confidence intervals, so our objective is different. If we wanted the confidence interval to have at least the target width and use the smallest sample size that would provide that width, we would use $n = 94$ and the width would be 8.039.

3.15.6 One-Sided Confidence Bounds

In some applications, a one-sided confidence bound is more appropriate than a two-sided confidence interval. Sample size can also be computed for one-sided confidence bounds (not "intervals"), although the software for accomplishing

this is somewhat limited. MINITAB has this capability for each type of two-sided confidence interval for which sample size can be determined, with the user specifying either an upper bound or a lower bound. For example, referring to the example discussed in Section 3.12, for which the sample size for the two-sided interval was 16, if we use the same inputs but specify that the sample size is to be determined for a 95% one-sided confidence bound, MINITAB gives the sample size as $n = 11$. (Of course, the sample size is the same regardless of whether a lower bound or upper bound is specified.) Why is the sample size less than in the two-sample case? The expression for, say, a 95% lower bound is: Lower Bound $= \bar{X} - Z_\alpha \sigma/\sqrt{n}$, with the error of estimation, E, equal to $\bar{X}$ − Lower Bound. So, $E = Z_\alpha \sigma/\sqrt{n}$ and thus $n = Z_\alpha^2 \sigma^2/E^2 = (1.645)^2\ (2)^2/1^2 = 10.82$, so $n = 11$ would be used. The smaller sample size results from the fact that $Z_\alpha < Z_{\alpha/2}$, as is always the case.

Other software that can be used for one-sided confidence bounds includes PASS and nQuery. Although Power and Precision does give confidence interval results, there is no option to select a confidence interval rather than a hypothesis test; the user simply solves for sample size for a hypothesis test and the corresponding confidence interval is part of the output.

Among well-known applets, Lenth's applet can also be used to solve for sample size for two-sided confidence intervals, but that is for one mean or one proportion, so its capabilities for confidence intervals is very limited.

3.16 RELATIVE PRECISION

As discussed by, for example, Lohr (1999, p. 38), sometimes there is an objective to achieve a desired relative precision, with the latter defined for estimating a mean as

$$\left| \frac{\bar{y} - \mu}{\mu} \right|$$

Thus, relative to the maximum error of estimation discussed in Section 3.13, for which the relevant probability statement would be $P(|\bar{y} - \mu| \leq E) = 1 - \alpha$, the corresponding probability statement for relative precision is $P(|(\bar{y} - \mu)/\mu| \leq E) = 1 - \alpha$, or equivalently, $P(|\bar{y} - \mu| \leq \mu E) = 1 - \alpha$.

Sample size determination for a maximum error of estimation for μ was discussed and illustrated briefly in Section 2.4. The appropriate formula (not given in that section) is

$$Z_{\alpha/2}\, \sigma/\sqrt{n} \leq E$$

which leads to the sample size formula

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E} \right)^2$$

The sample size formula for relative precision is obtained by substituting μE for E, which produces

$$n = \left(\frac{Z_{\alpha/2}\sigma}{\mu E} \right)^2$$

Since the population coefficient of variation (CV) was defined in Section 3.1.1 as σ/μ, we can thus write the formula as

$$n = \left(\frac{Z_{\alpha/2}CV}{E} \right)^2$$

[Note that this is *not* comparable to the sample size expression using CV given by van Belle (1993) and discussed in Section 3.1.1 because there is no assumption of a hypothesis test in this section.]

Thus, what is presented here is similar to the presentation for sample size using relative precision given in Lohr (1999, p. 40), with the exception that the latter assumes a finite population size and gives the formula using a finite population correction factor (fpc), as discussed in Section 2.3, although the fpc used by Lohr (1999) is $(N - n)/N$ rather than $(N - n)/(N - 1)$ that was given in Section 2.3.

3.17 COMPUTING AIDS

In addition to the software and applets that have been discussed and illustrated in this chapter and Chapter 2, Altman (1980) gave a nomogram that links power to sample size in comparing the means for two independent samples. This was also illustrated in Newcombe (2001). Nomograms for use in diagnostic studies were also given by Carley, Dosman, Jones, and Harrison (2005).

3.18 SOFTWARE

PASS, in particular, was used to compute sample sizes for various types of tests that were covered in this chapter, and nQuery, Power and Precision, MINITAB, Stata, and SAS were also mentioned. PASS has by far the broadest capability of the software that is specifically for sample size determination. Users do have

a wide variety of options for the standard statistical tests given in this chapter, and it has capabilities for some tests that were not covered in this chapter, nor are they covered in succeeding chapters. The `sampsi` and `sampncti` commands in Stata can be used for sample size determination, with development of the `sampncti` command motivated by the fact that the `sampsi` command does not work particularly well for small sample sizes (see Harrison and Brady, 2004).

3.19 SUMMARY

Determining sample size to test one or two means is a very common application of sample size determination and these applications were covered in this chapter. It is important to recognize that the assumed power for hypothesis tests is almost certainly not the actual power, so confidence limits for power for a fixed sample size are important, at least a lower confidence limit. Similarly, if a researcher wants to specify power and view an interval on sample sizes to produce a desired power, this would be another option. The bottom line is that power is not "known" and this should be addressed in practical applications. Of course, that problem can be avoided by constructing a confidence interval rather than performing a hypothesis test, and there is much support in the literature for doing so, but this does not completely avoid problems caused by unknown parameter values.

The tests and confidence intervals presented in this chapter are based on a normal distribution. When there is pronounced nonnormality, a nonparametric approach (Chapter 10) can be used. Alternatively, if a particular nonnormal distribution is known to be an adequate population model, sample sizes can be determined for such distributions. For example, Ren, Chu, and Lai (2008) gave sample size and power computations for a left-truncated normal distribution. Problems may be encountered with the specification of a nonnormal distribution, however, and Singh, Singh, and Engelhardt (1999) discussed some problems with the lognormal distribution assumption and how they relate to sample size determination.

APPENDIX

Equation (3.3) can be derived using the same general approach that was used to derive Eq. (2.3). That is, we start with

$$P\left[(\bar{X}_1 - \bar{X}_2) > \Delta_0 + Z_\alpha \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}} \,\middle|\, \mu_1 - \mu_2 = \Delta\right] = .80$$

with Δ_0 and Δ being the hypothesized and assumed values of $\mu_1 - \mu_2$, respectively. Then,

$$P\left[(\bar{X}_1 - \bar{X}_2) - \Delta > \Delta_0 - \Delta + Z_\alpha \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}} \,\middle|\, \mu_1 - \mu_2 = \Delta\right] = .80$$

Dividing through by $\sqrt{(\sigma_1^2 + \sigma_2^2)/n}$ produces, under the assumption of normality,

$$1 - \Phi\left(Z_\alpha + \frac{\Delta_0 - \Delta}{\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)/n}}\right) = .80$$

$$\Phi\left(Z_\alpha + \frac{\Delta_0 - \Delta}{\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)/n}}\right) = .20$$

so

$$Z_\alpha + \frac{\Delta_0 - \Delta}{\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)/n}} = -Z_\beta$$

Solving this last equation gives

$$n = \frac{\left(\sigma_1^2 + \sigma_2^2\right)\left(Z_\alpha + Z_\beta\right)^2}{(\Delta - \Delta_0)^2}$$

REFERENCES

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

Beal, S. L. (1989). Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics*, **45**, 969–977.

Bogle, W. and Y. S. Hsu (2002). Sample size determination in comparing two population variances with paired data: Application to bilirubin tests. *Biometrical Journal*, **44**, 594–602.

Boneau, C. A. (1960). The effects of violations of the assumptions underlying the *t*-test. *Psychological Bulletin*, **57**, 49–64.

Bonett, D. G. and M. S. Al-Sunduqchi (1994). Approximating sample size requirements for s^2 charts. *Journal of Applied Statistics*, **21**(5), 425–429

Bowman, K. O. and M. A. Kastenbaum (1974). Potential pitfalls of portable power. *Technometrics*, **16**(3), 349–353.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, **40**, 318–355.

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**(17), 1933–1940.

Campbell, M. J., S. A. Julious, and D. G. Altman (1995). Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *British Medical Journal*, **311**, 1145–1148. (Available at `www.bmj.com/content/311/7013/1145.full`.)

Carley, S., S. Dosman, S. R. Jones, and M. Harrison (2005). Simple nomograms to calculate sample size in diagnostic studies. *Journal of Emergency Medicine*, **22**, 180–181.

Clark, A. (1987). Approximate confidence bounds for estimated power in testing the general linear hypothesis. M. S. thesis, Department of Biostatistics, University of North Carolina, Chapel Hill.

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

Duncan, A. J. (1986). *Quality Control and Industrial Statistics*, 5th edition. Homewood, IL: Irwin.

Epstein, B. (1960). Statistical life test acceptance procedures. *Technometrics*, **2**(4), 435–446.

Evans, M., N. Hastings, J. B. Peacock, and C. Forbes (2010). *Statistical Distributions*, 4th edition. Hoboken, NJ: Wiley.

Falk, R. J., S. L. Hogan, K. E. Muller, and J. C. Jennette (1992). Treatment of progressive membrane glomerulopathy. *Annals of Internal Medicine*, **6**, 438–445.

Gibbons, C. (2000). *Determination of Power and Sample Size for Levene's Test*. Master's thesis submitted to The University of Colorado.

Greenwood, J. A. and M. M. Sandomire (1950). Sample size required for estimating the standard deviation as a per cent of its true value. *Journal of the American Statistical Association*, **45**(250), 257–260.

Grieve, A. P. (1991). Confidence intervals and sample sizes. *Biometrics*, **47**, 1597–1603.

Guenther, W. C. (1981). Sample size formulas for normal theory *T* tests. *The American Statistician*, **35**(4), 243–244.

Harris, M., D. J. Horvitz, and A. M. Mood (1948). On the determination of sample sizes in designing experiments. *Journal of the American Statistical Association*, **43**(243), 391–402.

Harrison, D. A. and A. R. Brady (2004). Sample size and power calculations using the noncentral *t*-distribution. *The Stata Journal*, **4**(2), 142–153.

Hartung, J. and G. Knapp (2010). Adaptive confidence intervals of desired length and power for normal means. *Journal of Statistical Planning and Inference*, **140**, 3317–3325.

Heasman, P. A., I. D. M. MacGregor, Z. Wilson, and P. J. Kelly (1998). Toothbrushing forces in children with fixed orthodontic appliances. *British Journal of Orthodontics*, **27**, 270–272.

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.

Kelley, K. and J. R. Rausch (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, **11**(4), 363–385.

Kieser, M. and G. Wassmer (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical Journal*, **38**(8), 941–949.

Kohavi, R. and R. Longbotham (2007). Online experiments: Lessons learned. *IEEE Computer*, 40(9), 103–105. (Available at `http://exp-platform.com/Documents/IEEEComputer2007OnlineExperiments.pdf`.)

Kohavi, R., R. M. Henne, and D. Sommerfield (2007). Practical guide to controlled experiments on the Web: Listen to your customers not to the hippo. Paper presented at KDD2007.

Kohavi, R., R. Longbotham, D. Sommerfield, and R. M. Henne (2009). Controlled experiments on the Web: Survey and practical guide. *Data Mining and Knowledge Discovery*, **18**, 140–181.

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

Lehmann, E. L. (1959). *Statistical Hypotheses*. New York: Wiley.

Lehr, R (1992). Sixteen *s* squared over *d* squared: A relation for crude sample size estimates. *Statistics in Medicine*, **11**, 1099–1102.

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

Lohr, S. L. (1999). *Sampling Design and Analysis*. Pacific Grove, CA: Duxbury. (The current edition is the 2nd edition.)

Mahnken, J. D. (2009). Power and sample size calculations for models from unknown distributions. *Statistics in Biopharmaceutical Research*, **1**(3), 328–336.

Markowski, C. A. and E. P. Markowski (1990). Conditions for the effectiveness of a preliminary test of variance. *The American Statistician*, **44**(4), 322–326.

Moher, D., C. S. Dulberg, and G. A. Wells (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association*, **272**, 122–124.

Nayak, B. K. (2010). Understanding the relevance of sample size calculation. *Indian Journal of Ophthalmology*, **58**, 469–470.

Nelson, P. R., P. S. Wludyka, and K. A. F. Copeland (2005). *The Analysis of Means* : *A Graphical Method for Comparing Means, Rates and Proportions*. ASA-SIAM Series on Statistics and Applied Probability. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Newcombe, R. G. (2001). Statistical applications in orthodontics, part III: How large a study is needed? *Journal of Orthodontics*, **28**(2), 169–172.

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

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, **31**(1,2), 9–12.

Rasch, D., K. D. Kubinger, and K. Moder (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical Papers*, **52**, 219–231.

Ren, S., H. Chu, and S. Lai (2008). Sample size and power calculations for left-truncated normal distribution. *Communications in Statistics—Theory and Methods*, **37**(6), 847–860.

Schouten, H. J. A. (1999). Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Statistics in Medicine*, **18**, 87–91.

Seelbinder, B. M. (1953). On Stein's two-stage sampling scheme. *Annals of Mathematical Statistics*, **24**, 640–649.

Singh, A. K., A. Singh, and M. Engelhardt (1999). Some practical aspects of sample size and power computations for estimating the mean of positively skewed distributions in environmental applications. United States Environmental Protection Agency, Office of Research and Development, Office of Solid Waste and Emergency Response, EPA/600/s-99/006.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 2**4**, 243–258.

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.

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

van Belle, G. and D. C. Martin (1993). Sample size as a function of coefficient of variation and ratio of means. *The American Statistician*, **47**(3), 165–167.

Wang, H., S.-C. Chow, and M. Chen (2005). A Bayesian approach on sample size calculation for comparing means. *Journal of Biopharmaceutical Statistics*, **15**, 799–807.

Wheeler, R. E. (1974). Portable power. *Technometrics*, **16**, 193–201.

Wheeler, R. E. (1975). The validity of portable power. *Technometrics*, **17**, 177–179.

Wludyka, P. S., P. R. Nelson, and P. R. Silva (2001). Power curves for the analysis of means for variances. *Journal of Quality Technology*, **33**(1), 60–65.

Wong, A. (2010). A note on confidence interval for the power of the one-sample t-test. *Journal of Probability and Statistics*. Open access journal, article ID 139856.

EXERCISES

3.1. Assume that $\sigma = 3$ and it is desired to detect with a one-tailed upper test, power of at least .80 when $\mu = 36$ and the null hypothesis is $\mu = 35$. Also assuming normality, what is the minimum sample size that will accomplish this? If the experimenter is set on using $n = 100$, what power will this provide?

3.2. Explain why it is not very logical to use an expression like Eq. (3.4) to solve for a sample size, even if the computation is performed by software but is based on Eq. (3.4).

3.3. Suppose we want to estimate the mean systolic blood pressure of men who are more than 25% above their ideal maximum body weight as indicated by body fat index charts. How many men should be sampled to estimate the mean with a 95% confidence interval of width 10 units? You will need to estimate σ from the range of blood pressures of men who are considerably overweight, perhaps searching the Internet for such information.

3.4. Kohavi, Longbotham, Sommerfield, and Henne (2009) gave a sample size determination example. An e-commerce site was assumed to exist such that 5% of the users who visited the site made some purchase, spending about \$75. The average amount spent over all visitors was thus assumed to be \$3.75, with the standard deviation assumed to be \$30. They stated that if a 5% change in revenue is to be detected with 80% power, over 409,000 site visitors would be needed. Although that might not be an exorbitant number for a popular website, they indicated that the number would be computed using the formula given in this chapter as Eq. (3.7).

(a) What is the actual number of site visitors that would be needed?

(b) If such a sample size is considered impractical, is there any action that the site owners might take to reduce the sample size and still have power of at least .80? Explain.

3.5. A clinical dietitian wants to compare the effectiveness of two different diets, A and B, for diabetic patients. She intends to obtain a random sample of diabetic patients from a very large group of patients and randomly assign them to each diet such that there will be an equal number of patients for each diet. The experiment will last 3 months, at the end of which a fasting blood glucose test will be administered to each patient. She doesn't know which diet will likely be shown to be the better one, but she is interested in detecting an average difference of 10 milligrams per deciliter between the diets, as she believes that is adequate for claiming superiority. In the absence of any prior information, she is willing to assume that the standard deviations for the two diets should be approximately equal, and from similar studies that she has read about in the literature, she estimates that the assumed common standard deviation should be about 15. She believes that the two populations should be approximately normal. How many observations should she have for each group if she uses a significance level of .05 and she wants the power to be .90?

3.6. Data from the now-famous Framingham study (`www.framinghamheartstudy.org`) permit a comparison of initial serum cholesterol levels for two populations of males: those who go on to develop heart disease and those who do not. The mean serum cholesterol level of men who did *not* develop heart disease is $\mu = 219$ milligrams per deciliter (mg/dL), with a standard deviation of 41 mg/dL. Assume that you wish to conduct a study of those men who *did* develop heart disease and you believe that their serum cholesterol level must have been at least 20 milligrams higher, and this is the minimum difference that you want to detect. You are willing to assume that the standard deviation for this group is approximately the same as for the other group. If $\alpha = .05$ and the power is to be .90, how many men should be in the study of men who did develop heart disease?

3.7. Consider the example in Section 3.6 with $\alpha = .05$, power $= .80$, $\Delta = 1$ and the assumption that $\sigma_1 = \sigma_2 = 1$ for a one-sided test of H_0: $\mu_1 = \mu_2$. For that example the solution was $n = 13$ for each group. Now assume, as in Example 3.4, that we want to have $n_2/n_1 = 2$. What sample sizes should be used? Note that $n_1 + n_2$ differs noticeably from $2(13) = 26$. Comment.

3.8. If the power for detecting a difference of 3.0 between two population means is .80, will the power be less than .80 or greater than .80 for detecting an actual difference of 2.0? Explain.