CHAPTER 4

Proportions and Rates

The need frequently arises to test the equality of two or more proportions, in addition to working with rates.

For example, the proportion of nonconforming units of a new, and purported improved manufacturing process might be compared with the proportion conforming under the standard process. The sample size could be determined so as to detect a certain minimum level of improvement.

The parameter λ of the Poisson distribution is frequently a rate, such as the rate at which some event occurs during a specified time period, such as the rate of arrivals of United Parcel Service trucks per hour at a particular UPS loading dock. There is also frequently interest in comparing two rates, such as incidence rates (e.g., exposure vs. no exposure to a toxin).

We will first consider proportions, noting that the software PASS lists a staggering number (81) of procedures that it has for proportions (one proportion, two proportions; independent proportions, correlated proportions; equivalence tests, noninferiority tests, etc). Many of these procedures are discussed and illustrated throughout the chapter.

4.1 ONE PROPORTION

We will first consider the case of a single proportion, with p_0 denoting the value of p under the null hypothesis, and p_1 denoting the true value. For a one-sided test, the usual textbook test statistic (but not necessarily the best approach), which is based on the assumed adequacy of the normal approximation to the binomial distribution is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \tag{4.1}$$

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

The test statistic is appropriate when both $\hat{p}$ and p_0 do not differ greatly from 0.5 when n is large and are close to 0.5 when n is not large. When this requirement is not met, an arcsin transformation should be used. One form of the transformation is $p^* = 2 \arcsin \sqrt{\hat{p}}$. The reason for the "2" is that $\text{Var}(p^*) = 1/n$, which is easier to work with than $1/4n$. When this transformation is used, it can easily be shown that the required sample size for a two-sided test with significance level α and power $1 - \beta$ is

$$n = \left(\frac{z_\beta + z_{\alpha/2}}{\arcsin(\sqrt{p_0}) - \arcsin(\sqrt{p_1})} \right)^2 \tag{4.2}$$

with p_0 and p_1 denoting the null hypothesis value of the population proportion and the value that the experimenter wishes to detect, respectively. As with sample size formulas in previous chapters, $z_{\alpha/2}$ would be replaced by z_α for a one-sided test. Note that the use of the z terms means that a normal approximation is being used; the proportions are simply being transformed before the normal approximation is applied.

To illustrate, if $p_0 = .2$, $p_1 = .1$, and a one-sided test is to be performed with $\alpha = .05$ and power $= .90$,

$$\begin{aligned} n &= \left(\frac{z_\beta + z_\alpha}{2\arcsin\left(\sqrt{p_0}\right) - 2\arcsin\left(\sqrt{p_1}\right)} \right)^2 \\ &= \left(\frac{1.28155 + 1.64485}{0.927295 - 0.643502} \right)^2 \\ &= 106.332 \end{aligned}$$

so $n = 107$ would be used. One software package that uses this transformation in determining sample size is Power and Precision, as the only other option is to use the exact binomial distribution approach. Power and Precision does give $n = 107$ as the sample size for this example.

This transformation will often be needed because, for example, if we are looking at performance measures of a company measured by proportions, we would expect proportions to be not far from 0 or 1, depending on whether what is being measured is good or bad, and thus a proportion near 0.5 could be uncommon.

The transformation can also be used when there are two proportions; this is illustrated and discussed in Section 4.2.

Of course, a major disadvantage of transformations is that a transformed proportion generally won't have any practical meaning. For confidence intervals, there is some evidence (Vollset, 1993; Newcombe, 1998) that the Score method with continuity correction may provide the best results when p is very small ($\leq .01$).

Another option is to use the binomial distribution directly, and to determine sample size for confidence intervals using the "exact" approach, which the software PASS terms the "Exact (Clopper–Pearson)" option, following Clopper and Pearson (1934). Samuels and Lu (1992) provided guidelines, expressed as a percentage of the length of the Clopper–Pearson interval, for when the normal approximation confidence interval $\left(\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right)$ is an adequate substitute for the Clopper–Pearson interval. Tobi, van den Berg, and de Jong-van den Berg (2005) concluded from their study that the best method to use in reporting confidence intervals when p is less than or equal to .01 is to use either the Exact method or the Score method with continuity correction.

To illustrate the Exact method, Clopper and Pearson (1934) gave an example for which a crude measure of quality for a particular manufacturing process has been the percentage of articles that must pass a certain test. The percentage has fluctuated around $p = .60$ but an intensive effort is to be undertaken to improve quality and thus increase the percentage. After completion of that effort, a sample is to be undertaken to see if there is evidence of improvement and the objective is to determine how large a sample should be obtained such that a 95% confidence interval for p will have a width of .05. Clopper and Pearson (1934) gave confidence belts for p for various sample sizes, but those figures could not be used to solve for the exact, or even approximate, value of n. So the best that they could do was to say that the sample size must exceed 1000, based on .60 as an estimate for p. The sample size can be obtained using PASS and selecting the "Exact (Clopper–Pearson)" confidence interval method in the "Confidence Intervals for One Proportion" procedure. Doing so produces $n = 1513$, which is indeed (much) greater than 1000. This solution can easily be checked, if desired, by using software that gives binomial distribution probabilities. For example, using MINITAB, the (approximate) 97.5 percentile of the binomial distribution when $p = .60$ and $n = 1513$ is $X = 945$, with X being the binomial random variable. The approximate 2.5 percentile is 870. Since $870/1513 = 0.575$ and $945/1513 = 0.625$, to three decimal places, the confidence interval about 0.60 thus has width 0.05, as desired.

One objection to this approach might be that the necessary sample size is determined assuming that the proportion is being estimated without error. Gould (1983) gave a method for determining sample size for a proportion that incorporates estimation uncertainty explicitly in the form of joint confidence distributions obtained from a pilot study or from prior information.

When the test statistic given in Eq. (4.1) is used, *one possible* expression for n (derived in the chapter Appendix, but note the discussion there, especially in regard to software) is

$$n = \left[\frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p'(1-p')}}{p' - p_0}\right]^2 \tag{4.3}$$

■ EXAMPLE 4.1

To illustrate, assume a one-sided test and that $p_0 = .5$, $p' = .6$, the desired value of β is .80, and $\alpha = .05$. Then

$$n = \left[\frac{1.64485\sqrt{.5(1-.5)} + 0.841621\sqrt{.6(1-.6)}}{.6-.5}\right]^2$$

$$= 152.457$$

so n is set equal to 153, with the resultant power then being slightly greater than .80 since the rounding is upward. It should be noted that whereas this is the solution obtained using MINITAB, Power and Precision, and nQuery, PASS uses this normal approximation approach as its starting point for the purpose of obtaining the critical value of the test statistic, but then uses a search procedure starting from a sample size of one and obtains $n = 146$. The power is .8059 and since this exceeds the target power of .80, one could contend that this is a better solution than $n = 153$ when the cost of sampling is considered.

If the test had been two-sided, the only change to Eq. (4.3) is that z_α would be replaced by $z_{\alpha/2}$. For this example, but with the test two-sided instead of one-sided with $z_{\alpha/2} = 1.96$, the sample size would be computed as $n = 193.852$ and rounded up to 194.

As in any sample size determination problem, it is desirable to conduct a sensitivity analysis, to see how power changes as the sample size changes. For a two-sided test of $p_0 = .5$ when the true value is $p' = .6$, the relationship between power and sample size is shown in Figure 4.1.

The curvilinear pattern is obvious, and a regression model fit with linear terms in n and n^2 provides almost an exact fit, as the sum of the squared residuals is 0.0002. Thus, the regression model $\hat{Y} = 0.046796 + 0.00567292\ n - 0.00000921\ n^2$ produces the power values almost exactly. (Such a fitted equation could be useful if there is some indecision or disagreement about the sample size to use in a particular application, as possible sample sizes and the corresponding power values could easily be compared.) From a practical standpoint, the convexity of the curve means that there are (slightly) diminishing increases in power for constant increases in the sample size for the range of sample sizes depicted in Figure 4.1.

Although this approach will often be satisfactory, an approximation based on the beta distribution will often be preferred, and an exact test based on the binomial distribution will frequently be better than both approximations (as we would guess), although exact tests do have some shortcomings. In particular, they have been criticized as being too conservative. Nevertheless, exact test results will be discussed later.

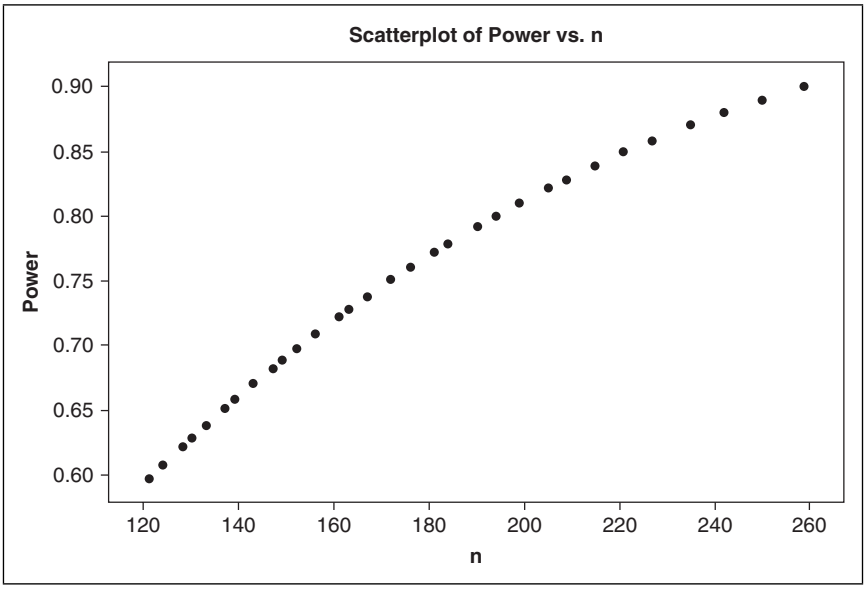

Figure 4.1 Power versus sample size for two-sided test: $p_0 = .5$, $p' = .6$, $\alpha = .05$.

As stated previously, there is nothing magical about a power value of .80. If more observations than are needed to give a power of .80 can be afforded and power of greater than .80 is desired, graphs such as Figure 4.1 can be helpful.

If the test had been one-sided, as in the original example, then Figure 4.2 shows the relationship between power and sample size.

As in Figure 4.1, the curvilinear relationship between power and sample size is obvious, and as before, a regression model with linear and quadratic terms in sample size provides an almost exact fit.

Since the true value of p is, of course, unknown, it may also be of interest to look at the sample size requirement for other values of p', such as .55, .65, and .70 or some subset of these numbers. A single graph using multiple values of p' can be constructed using software such as Power and Precision. ■

4.1.1 One Proportion—With Continuity Correction

Although Eq. (4.3) is commonly used to determine sample size, Fleiss, Levin, and Paik (2003) recommended using the continuity correction $n^* = n + 1/\delta$, with n denoting the sample size before the correction and $\delta = z_\alpha\sqrt{p_0\,(1 - p_0)/n} + z_\beta\sqrt{p_1(1 - p_1)/n}$. For a two-sided test, z_α would be replaced by $z_{\alpha/2}$.

If we apply this to Example 4.1, we obtain $n^* = 155 + 1/.0998 = 165$. Thus, use of the continuity correction factor makes a noticeable difference, although use of the continuity correction for determining sample size is somewhat controversial,

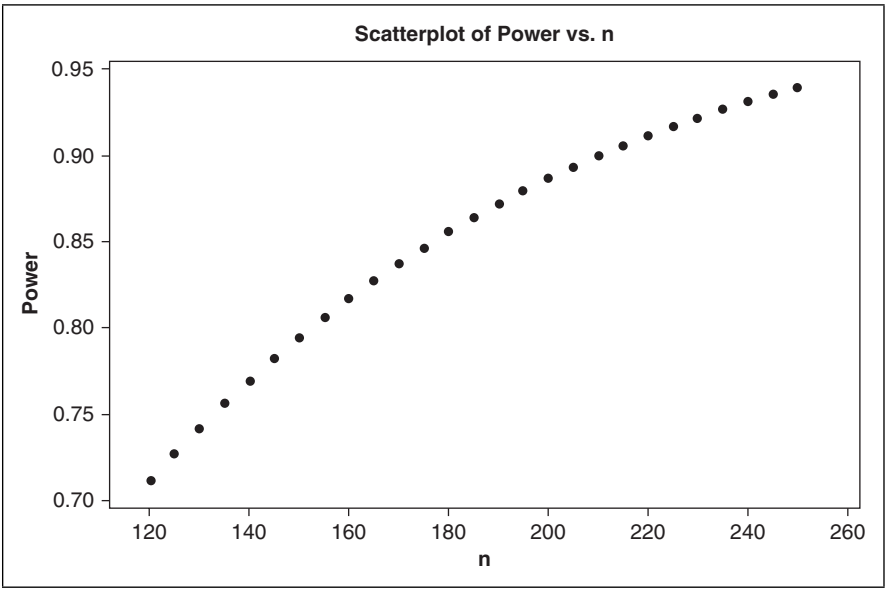

Figure 4.2 Power versus sample size for one-sided test: $p_0 = .5$, $p' = .6$, $\alpha = .05$.

as Gordon and Watson (1996) recommended that it *not* be used for inferences involving two proportions.

The corresponding hypothesis test with the continuity correction is

$$Z = \frac{\hat{p} - p_0 - c}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

with the value of c dependent on the form of the alternative hypothesis and thus essentially dependent on the sign of $(\hat{p} - p_0)$. If the sign is positive, then $c = 1/2n$; if the sign is negative, $c = -1/2n$; with no adjustment made if $|\hat{p} - p_0| < 1/2n$. The logic behind these adjustments is the same as the logic behind the $\pm 1/2$ adjustment that is used when $X = n\hat{p}$ is used instead of $\hat{p}$. That is, since X is discrete and we are using a normal approximation (i.e., continuous), X is converted to the continuum $(X - 1/2, X + 1/2)$. It then follows that $X/n = \hat{p}$ would be converted to the continuum $(X/n - 1/2n, X/n + 1/2n)$.

4.1.2 Software Disagreement and Rectification

There is the potential for confusion when users try different sample size determination software and receive different answers for the same problem. This can especially be a problem involving proportions, as the reader should note. To

illustrate, Chow, Shao, and Wand (2008, p. 87) gave an example of testing a single proportion, for which the null hypothesis is $p_0 = .3$, but after a particular treatment the proportion is expected to be around .5. So .5 is to be detected with power .80 and $\alpha = .05$ is used for a two-sided test. In deriving their expression for the power of the test using a normal approximation (i.e., using Z), the authors stated: "When n is large, the power of the test is approximately . . . ." Thus, they did not derive the exact expression for the power of the test, which means that their expression for n is not exact for a normal approximation. We will see that their solution, $n = 49$, does not agree with the solution for *any* of the software discussed so far. If Eq. (4.3) is used, $n = 44$ is obtained, which is also the solution given by nQuery, SiZ, and Lenth's applet. MINITAB, PASS, and Power and Precision give $n = 47$, however. Why the difference? As stated in Section 4.1, Power and Precision uses an arcsin transformation of p, so the use of Eq. (4.1) should produce $n = 47$. Thus,

$$
\begin{aligned}
n &= \left(\frac{z_\beta + z_\alpha}{2\text{arcsin}\left(\sqrt{p_0}\right) - 2\text{arcsin}\left(\sqrt{p_1}\right)} \right)^2 \\
&= \left(\frac{0.84 + 1.645}{2\text{arcsin}\left(\sqrt{.3}\right) - 2\text{arcsin}\left(\sqrt{.5}\right)} \right)^2 \\
&= 46.2957
\end{aligned}
$$

so $n = 47$ would be used, in agreement with MINITAB, PASS, and Power and Precision.

Thus, although these software packages are using a normal approximation, there is an attempt to first normalize the sample proportion since a normal approximation does not do so. That is desirable when p is small and much less than .5, although here it is questionable whether or not an arcsin transformation is necessary. See also the discussion of an arcsin transformation for proportions data in Section 4.2.

4.1.3 Equivalence Tests and Noninferiority Tests for One Proportion

Recall that equivalence, noninferiority, and superiority testing were discussed briefly in Section 1.5. For proportions data, a single proportion might be considered "equivalent" to the hypothesized value if the difference is small enough to be deemed inconsequential.

Of the software mentioned previously, PASS is the only software that can be used for sample size determination for equivalence tests with a single proportion. To illustrate, let power = .80, $\alpha = .05$ for a two-sided test, and assume that $p =$ .60 has been the standard proportion value and that this is expected to continue

(i.e., p and p_0 are the same). We will use 0.2 as the equivalence value, meaning that all values in the interval (0.40, 0.80) are considered to be equally good. Chow et al. (2008, p. 88) gave $n = 52$. They used a large sample approximation in deriving the expression for power from whence came their expression for n, which is

$$n = \frac{\left(z_\alpha + z_{\beta/2}\right)^2 p\,(1-p)}{(\delta - |p - p_0|)^2}$$

Thus, for this problem,

$$\begin{aligned} n &= \frac{\left(z_\alpha + z_{\beta/2}\right)^2 p\,(1-p)}{(\delta - |p - p_0|)^2} \\ &= \frac{(1.645 + 1.28)^2(.6)(.4)}{(.2 - |.6 - .6|)^2} \\ &= 51.337 \end{aligned}$$

which is then rounded up to $n = 52$.

Notice that this solution is obtained despite the fact that $p = p_0$; that is, the true proportion and hypothesized proportion are assumed to be the same. Would this equality work when PASS is used? Yes, provided that the "Equivalence Tests for One Proportion (Differences)" is used. They give several options for the form of the test statistic and although their solutions do not agree with the Chow et al. (2008) solution, for some options the solution is $n = 53$, which differs by only one unit from the solution given by Chow et al. (2008). One source of difference is that the latter is based on hitting the desired power and significance levels exactly, which generally will not be possible with discrete data and distributions, with the PASS output giving the actual power and significance level, which differ somewhat from the target values. (The formula or approach that PASS uses is not indicated.) See also the discussion for sample size determination for equivalence testing of proportions given by Farrington and Manning (1990).

4.1.4 Confidence Interval and Error of Estimation

Since $\sigma_p = \sqrt{p(1-p)/n}$, which is thus a function of the parameter that is to be estimated, it is not possible to obtain an exact expression for the sample size so as to give a confidence interval for p of a specified width since the (large-sample) $100(1-\alpha)\%$ confidence interval is of the form

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \tag{4.4}$$

if a continuity correction is not used, with the width of the interval being $2z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$. If a continuity correction *is* used, it would be reasonable for the form of the confidence interval to correspond to the form of the hypothesis test when the continuity correction is used. This leads to the upper limit (*U.L.*) and lower limit (*L.L.*) given by

$$U.L. = \frac{\hat{p} + \dfrac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n} + \dfrac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}$$

$$L.L. = \frac{\hat{p} + \dfrac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n} + \dfrac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}$$

This form of the interval has been supported by various researchers, including Agresti and Coull (1998). See Ryan (2007, p. 159) for further discussion.

Like the hypothesis test, this approach is based on the assumed adequacy of the normal approximation to the binomial distribution, which will not necessarily work when p is small. [The value of p is more important than the products np and $n(1-p)$ in determining whether or not the normal approximation should be adequate. See, for example, Schader and Schmid (1989).]

The maximum error of estimation of p is half the width of the confidence interval, with $\hat{p}$ being in the center of the interval. Specifically, since the width of the interval is Upper Limit $-$ Lower Limit $= \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} - (\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}) = 2z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$), the halfwidth is thus $z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$.

Since the halfwidth is the maximum error of estimation with probability α, we could set $z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = E$ and solve for n, after we substitute a value for $\hat{p}$. Before doing the latter, we have

$$n = \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{E^2}$$

Of course, we don't have a value for $\hat{p}$ before a sample has been taken, and whatever value is substituted for $\hat{p}$ in solving for n would almost certainly not be the value of $\hat{p}$ that results when the sample is taken. One approach that has been used is to substitute 0.5 for $\hat{p}$, as this maximizes $\sigma_{\hat{p}}$, and thus maximizes the value of n that is obtained when solving for it for a specified confidence interval width.

Doing so gives $n = Z^2_{\alpha/2}/4E^2$. Thus, for example, if $E = .05$ and a 95% confidence interval is to be used, $n = 100(1.96)^2 = 384.2$, so $n = 385$ would be used. (This value for E was selected deliberately, as it is useful for Example 4.2.)

This is essentially an infinite population size formula, as it is not a function of the population size. Furthermore, one could argue that relative error is more important than absolute error when p is very small. An alternative formula, which is a function of both the population size, N, and the relative error, ϵ, was given by Levy and Lemeshow (1991). For a 95% confidence interval, the expression is

$$n = \frac{1.96^2 Np(1-p)}{(N-1)\epsilon^2 p^2 + 1.96^2 p(1-p)}$$

If the objective were to estimate the population total rather than the population proportion, the total, Np, would be estimated by $N\hat{p}$, but the applicable sample size formula would *not* be simply $N^2\left(Z^2_{\alpha/2}/4E^2\right)$ if absolute rather than relative error were used because E would obviously have to be redefined since a total would be estimated with greater absolute error than a proportion. Therefore, the sample size expression would be $n = N^2\left[Z^2_{\alpha/2}/4(E^*)^2\right]$ for a suitably chosen value of E^*, if the finite population correction factor, $1 - n/N$, were ignored. It would be quite reasonable to let $E^* = NE$ since the total is estimated by $N\hat{p}$ and E is the acceptable error for estimating p.

■ EXAMPLE 4.2

Motivated by his work as an expert witness in one particular court case, Afshartous (2008) proposed a new way of looking at sample size determination for determining a confidence interval for a binomial proportion, with the proportion of interest being the proportion of bills sent out by a company that included a "special service handling fee," with such bills alleged to be erroneous. The Defendant proposed a margin of error of .05, which of course wouldn't make any sense if the proportion were close to zero. In particular, the computed lower confidence limit could be negative, so there would then not be a lower limit.

Afshartous (2008) presented the sample size formula in an awkward way because it was given as $n = (1.96\,Ns)^2/E^2$, with $s^2 = s_y^2$, the sample variance for the realization of each of the n Bernoulli trials that form the binomial experiment and this E different from the E used previously and not necessarily equal to E^*.

It is somewhat awkward to give a sample size formula in terms of a sample variance since the sample won't be taken until the sample size is determined. Afshartous (2008) stated that a value would have to be used for s, but selecting a value may be difficult and of course would just be an approximation. In arguing for the use of relative error, Afshartous (2008) stated that it would be difficult to specify a value for E since the total is unknown. Of course, the use of $E^* = NE$ would be reasonable if E were reasonable. As indicated previously, however, this won't work if p is small.

Because of these difficulties, Afshartous (2008) contended that it would be better to solve for n using the coefficient of variation. The population coefficient of variation for a random variable Y is defined as σ_Y/μ_Y, with the corresponding sample coefficient of variation given by $s_Y/\bar{y}$. Since the expected value of $\hat{p}$ equals p, the expected value of $N\hat{p}$ is Np. Since $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, $\sigma_{N\hat{p}} = N\sqrt{p(1-p)/n}$. Then, $\sigma_{N\hat{p}}/\mu_{N\hat{p}} = (1/\sqrt{n})\sqrt{(1-p)/p}$.

This last expression could be set equal to a value for the coefficient of variation and n solved for, but selecting a desired value for the coefficient of variation is likely to be more difficult than selecting a value for either E or E^*. Since the Plaintiff's lawyers had trouble just understanding the concept of a "coefficient of variation," Afshartous instead used "margin of error," with this term obviously used to represent the halfwidth of a confidence interval. The author stated that the standard deviation is roughly one-half the margin of error. (This is obviously based on the use of $z = 1.96$ for a 95% confidence interval.) This led to Afshartous (2008) defining Accuracy = Margin of Error/True Parameter, using the author's notation. This could be written as $(1/\sqrt{n})\sqrt{(1-p)/p}$= Accuracy/2, with a set of possible values of n then obtained by solving for n using various reasonable combinations of Accuracy and p. This in turn led to Table 2 in Afshartous (2008), which enabled the lawyers to better understand the results. The eventual sample size that was used was stated as being "rather large" but was not given for obvious reasons of confidentiality. The margin of error for the total was 12.2%, with the sampling done so as to provide a confidence interval to be used in mediation.

The "moral of the story," if there is one, is that everyone involved in the settling of a legal dispute needs to understand the statistical methodology that is being used. ■

4.1.5 One Proportion—Exact Approach

Although most software assume the use of the normal approximation to the binomial distribution, without a continuity correction, some software give the user the option of using an exact approach: that is, using the probabilities for the binomial distribution, which has considerable intuitive appeal. To illustrate, consider again Example 4.1. Using PASS, we find the exact solution to be $n = 158$.

Exact tests are often computer intensive and are not easy to illustrate because the solution results from the use of a search algorithm rather than the solution of an equation. That is the case here but we can still gain insight into the solution, if not see how it is obtained. If we use software to produce the cumulative mass function for the binomial distribution for $n = 158$ and $p = .5$, we see that $P(X \geq 90) = .04724$, and this is the closest we can come to .05. Therefore, the null hypothesis will be rejected when $X \geq 90$. Power is then computed as $P(X \geq 90|p = .6)$. Using MINITAB or other software, we can see that $P(X \geq 90|p = .6) = .80565$.

The output from PASS is given below, rounded to four decimal places.

```
                    Power Analysis of One Proportion
Page/Date/Time        1  10/20/2009 5:52:59 AM

Numeric Results for testing H0: P = P0 versus H1: P > P0
Test Statistic: Exact Test

                Proportion  Proportion                      Reject  H0
                Given H0    Given H1    Target  Actual              If
Power      N    (P0)        (P1)        Alpha   Alpha   Beta    R>=This
0.8057   158    0.5000      0.6000      0.0500  0.0472  0.1943  90
```

Chernick and Liu (2002) illustrated a quirk of the exact approach as power can actually decrease as the sample size is increased. Thus, the relationship between power and sample size is not monotonic. This is not a software problem: whereas they obtained such nonintuitive results using nQuery Advisor, the same results are obtained using PASS and presumably any other software that has capability for the exact approach. For example, using the Chernick and Liu (2002) example of performing a one-sided test with $\alpha = .025$ of $p_0 = .07$ and assuming that $p_1 = .03$ is the true value, PASS gives the power as .8126 when $n = 240$ and .7995 when $n = 244$. When PASS is used to solve for n with the target power specified as .80, it gives the solution as 240, whereas nQuery Advisor does not permit sample size to be solved for with their exact test routine. Instead, n is entered and the value of power is produced. Of course, this could require at least a modest amount of trial and error if the user has no idea how large the sample size should be in a particular application in order to produce the desired power. Their sidebar states that power does not increase monotonically with sample size and refers to Chernick and Liu (2002).

Chernick and Liu (2002) showed that in general there is a sawtooth pattern when power is graphed against sample size. Why does this happen and should it be of any concern? These types of problems should be expected whenever a discrete random variable is involved, as noted by Cesana, Reina, and Marubini (2001), who also noted the sawtooth pattern in presenting their two-step procedure for determining sample size for testing a proportion against a reference value. The values of the binomial random variable that result in rejection of the null hypothesis are 9 for both $n = 240$ and $n = 244$. Since this is a lower-tailed test, the cumulative probability at that reject number for each sample size is of interest. Since the reject number does not increase as the sample size increases, there is less cumulative probability at that point with the larger sample size because the total probability of one is spread over a larger number of possible values of the random variable (245 versus 241). Therefore, with the reject number being the same, the power must be smaller with the larger sample size. Thus, any computer algorithm that seeks to come as close as possible to the target power but does not require that the target power be exceeded is going to exhibit this quirk.

Although not discussed by Chernick and Liu (2002), the significance level also jumps around for these two sample sizes, being the specified value of .025 for $n = 240$, but being considerably less at .0214 for 244.

Chernick and Liu (2002) raised the question of which sample size should be used since although the power is greater than .80 at $n = 240$ and that is the smallest sample size for which that occurs, the power dips below the required .80 at 244. The authors reported that a colleague suggested that $n = 277$ be used, as this is the smallest sample size for which the power exceeds .80 at all larger sample sizes. It seems unwise to think that way, however, because parameter values assumed in the null hypothesis aren't going to be equal to the hypothesized value, so this isn't truly going to be an "exact" calculation (i.e., null hypotheses are almost always false, as was discussed in Section 1.1). What if p_0 had been .072 instead of .070? Then the necessary sample size would have been 215, which is quite different from 240. Sample size determination is never going to be an exact science, so being overly concerned with "exactness" in certain ways seems inappropriate.

4.1.6 Bayesian Approaches

M'Lan, Joseph, and Wolfson (2008) discussed estimating binomial proportions using Bayesian techniques and provided tables for determining sample size for each of several methods that they presented. Their paper is long, detailed, and at an advanced level and a discussion of it would be beyond the level of this book. Readers who are interested in using Bayesian methods in estimating binomial proportions may find the paper useful, however. In earlier work by two of these three authors, Joseph, Wolfson, and du Berger (1995) considered the use of Bayesian approaches for binomial data and for higher-dimensional applications. Bayesian sample size determination with a focus on estimating the binomial parameter was also considered by Pham-Gia and Turkkan (1992) and since Bayesian approaches have not generally been incorporated into sample size determination software, it is worth noting that, even though this was two decades ago, the authors stated a computer program that "handles all computational complexities" was available upon request.

4.2 TWO PROPORTIONS

Often it is desirable to test for the equality of two proportions: are two competing drugs equally effective, do two surgical procedures give equivalent results, and so on. There are several ways of doing so, however, and the selection is a bit complex and not textbook-simple. The determination of the sample size(s) for the two populations will depend on which test method is used. Some software offer several options, but some leading software offers very few options.

Thus, H_0: $p_1 = p_2$ is tested, which of course is the same as testing $p_1 - p_2 = 0$. *One* way to test this hypothesis is to use the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\dfrac{2\bar{p}(1-\bar{p})}{n}}} \tag{4.5}$$

with $\bar{p} = (\hat{p}_1 + \hat{p}_2)/2$ assuming that $n_1 = n_2 = n$. This test statistic assumes the adequacy of the normal approximation to the binomial distribution and will work well when $\hat{p}_1$ and $\hat{p}_2$ are both close to .5. This will often not be the case, as in many applications, such as the percentage of people who have a reaction to a drug or the percentage of nonconforming units in a manufacturing process, the proportions will be much less than .5. Then it is necessary to apply an arcsin transformation to the proportions or to use software that will perform the transformation and then determine the required sample size, such as the Power and Precision software.

Although an arcsin transformation is typically applied to data, it could also be applied directly to proportions. As stated in Section 4.1, one form of the transformation is $2 \sin^{-1}\sqrt{p}$ and the transformed proportion would then be used in place of each proportion. Although the transformation should be used before a normal approximation approach is employed when proportions are small, it can make only a small percentage difference at times in the computed sample size.

To illustrate, assume that we wish to test the equality of two proportions but for the purpose of sample size determination we want to be able to reject the null hypothesis of equal proportions when $p_1 = .01$ and $p_2 = .02$. With $\alpha = .05$, power $= .80$, and a one-sided test, $n = 1826$ for each sample if the normal approximation *without* the arcsin transformation is used, and $n = 1776$ when the arcsin transformation is used, as given by Power and Precision. (The sample size formula using a normal approximation with no arcsin transformation is given later in this section.) The sample size for a one-sided test using the arcsin transformation is computed as

$$\begin{aligned} n &= \frac{1}{2}\left(\frac{z_\beta + z_\alpha}{\arcsin(\sqrt{p_1}) - \arcsin(\sqrt{p_2})}\right)^2 \\ &= \frac{1}{2}\left(\frac{0.84 + 1.645}{\arcsin(\sqrt{.01}) - \arcsin(\sqrt{.02})}\right)^2 \\ &= \frac{1}{2}\left(\frac{0.841621 + 1.64485}{0.100167 - 0.141897}\right)^2 \\ &= 1775.17 \end{aligned}$$

so that $n = 1776$ would be used, in agreement with Power and Precision.

Thus, the percentage difference between the two sample sizes is small. This is because the sample sizes are so large, which results from the very small difference between the two population proportions. Now let $p_2 = .10$, so that the difference in the two proportions is not small. Now $n = 63$ when the arcsin transformation approach is used and $n = 79$ when the normal approximation approach is used. Now the percentage difference between the two sample sizes is not small, and that is because the percentage difference between the two proportions is not small. Here the arcsin transformation approach should definitely be used, not only because the proportions are small, which necessitates its use, but in this case it makes a sizable difference in the sample size.

In sample size determination, the objective in a test of equality would be to determine the (usually common) sample size such that a specified nonzero difference will be detected with a high probability. The formula for the common sample size using a normal approximation and assuming a one-sided test is

$$n = \left[\frac{z_\alpha \sqrt{(p_1 + p_2)(1 - p_1 + 1 - p_2)/2} + z_\beta \sqrt{p_1(1 - p_1) + p_2(1 - p_2)}}{p_1 - p_2} \right]^2 \tag{4.6}$$

(This formula is also derived in the chapter Appendix.) If the test is two-sided, z_α would be replaced by $z_{\alpha/2}$.

To illustrate, if a difference of .1 between the two proportions is to be detected with p_1 assumed to be .55 and p_2 assumed to be .65, with an alternative hypothesis of $p_1 < p_2$ and a probability of .80 using a one-sided test with a significance level of .05, the common sample size would be 296, which would give a power of .80, as the reader is asked to show in Exercise 4.1. PASS also gives 296 as the solution, as does nQuery, MINITAB, Lenth's applet, Power and Precision, and G*Power. Note the software agreement for the case of two proportions whereas there was considerable disagreement for one proportion because of the different methods that were used.

If the test had been two-sided, the result would have been $n = 375.14$ (rounded up to 376), which will be used later in this section for comparison purposes.

Although approximate sample size formulas aren't necessary since Eq. (4.4) is known and software are available, Campbell, Julious, and Altman (1995) gave an approximation formula that is simpler than Eq. (4.4). Specifically, for a power of .80 and a significance level of .05 for a two-tailed test, they gave $n = 16\bar{p}(1 - \bar{p})/(p_1 - p_2)^2$, with $\bar{p} = (p_1 + p_2)/2$, although they stated that this does slightly overestimate the sample size. (They did not discuss a one-tailed test.)

Another approximation was given by Fleiss, Tytun, and Ury (1980) and Fleiss, Levin, and Paik (2003, p. 73) which is, *in their notation,*

$$n = n' + \frac{2}{|P_2 - P_1|} \tag{4.7}$$

with their n' the same as the n in Eq. (4.6). They indicated that this should be used when $n'|P_2 - P_1| \geq 4$ and stated that Eq. (4.7) is useful both in quickly estimating required sample sizes and in estimating power. They stated that the corresponding power can be obtained by inverting Eq. (4.6) and Eq. (4.7) to obtain

$$z_\beta = \frac{|P_2 - P_1|\sqrt{n - \dfrac{2}{|P_2 - P_1|}} - z_{\alpha/2}\sqrt{2\bar{P}\bar{Q}}}{\sqrt{P_1Q_1 + P_2Q_2}}$$

with the power then easily obtainable from Z_β.

To illustrate the Campbell et al. (1995) approximation, if we assume, as in the previous example, that $p_1 = .55$ and $p_2 = .65$, with an alternative hypothesis of $p_1 < p_2$ and desired power of .80 using a one-sided test with a significance level of .05, their approximation gives a sample size of $n = 16\bar{p}(1 - \bar{p})/(p_1 - p_2)^2 = 16(.6)(.4)/(.55 - .65)^2 = 384$, which is, as expected, slightly larger than $n = 376$ obtained using Eq. (4.4). The solution using the Fleiss et al. (1980) approximation is $n = 396$, as can be seen from Eq. (4.7) combined with the solution from Eq. (4.6).

So which of the three solutions should a practitioner use? The question of sample size for testing the equality of two proportions will be addressed more broadly after considering the exact test discussed later in this section and the use of a continuity correction discussed in Section 4.2.1, but the touted "remarkable accuracy" (Fleiss et al., 2003, p. 73) of the Fleiss et al. (1980) approximation can be questioned, at least for this example, since the Campbell et al. (1995) approximation generally overestimates the necessary sample size but for this example the Fleiss et al. (1980) approximation gives an even larger sample size than does the Campbell et al. (1995) approximation!

Another commonly used test for testing the equality of two independent proportions against an inequality alternative is a chi-square test. It can be shown with the appropriate algebra that the chi-square test (using a 2×2 contingency table) without a continuity correction is equivalent to the Z-test with no continuity correction. (This result should not be surprising since the square of a standard normal random variable is a chi-square random variable with one degree of freedom.) See Suissa and Shuster (1985), who gave sample size methods for the chi-square test with the 2×2 contingency table, and compare these results with the sample sizes obtained using other tests.

We will label the entries in the two-way table and the marginal totals as follows.

	Sample 1	Sample 2	Totals
Success	a	b	s
Failure	c	d	f
Totals	m	n	N

The chi-square statistic for general usage is typically written as

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

with, in this case, 4 denoting the number of cells in the 2×2 table, O_i denotes the observed value in the ith cell and E_i denotes the corresponding expected value, computed under the assumption that the proportions are equal. It can be shown (see, e.g., Conover, 1980, p. 145) that for the 2×2 table the chi-square statistic can be written as

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \tag{4.8}$$

It can be shown by appropriately inserting these letters in Eq. (4.5) that the square of the Z-statistic in Eq. (4.5) is equal to χ^2 in Eq. (4.8), as the reader is asked to do in Exercise 4.6. This should not be surprising because a random variable that has a chi-square distribution results from squaring a random variable that has a standard normal distribution, whereas the use of a Z-statistic implies either normality or a normal approximation.

Thus, the same conclusion will be reached with each test, which implies that the sample size computed for each test will be the same.

We will examine the other options for testing two proportions in Sections 4.2.1 and 4.2.2 and then try to draw a conclusion about what approach should be used by practitioners.

4.2.1 Two Proportions—With Continuity Correction

Although the formula in Eq. (4.6) is commonly used, and is what is given in introductory level textbooks, it might seem preferable to use the formula that incorporates a continuity correction, just as in the single proportion case.

With two proportions there is the obvious question of how the continuity correction should be used and should there be a correction for each proportion. If the alternative hypothesis is $p_1 > p_2$, one might assume that using $\hat{p}_1 + 1/2n_1$ and $\hat{p}_2 + 1/2n_2$ in the test statistic would be appropriate because this would be

most favorable to that hypothesis (with the signs reversed for $p_1 < p_2$) but of course there should be justification for whatever continuity correction is used.

Levin and Chen (1999) examined the approximation of Casagrande, Pike, and Smith (1978a) and explained why it has "remarkable accuracy," which is because it is a valid approximation to the exact hypergeometric test procedure [as explained by Fleiss et al. (2003, p. 73)].

For a one-sided test, the sample size formula with the continuity correction is

$$n^* = \frac{n}{4}\left(1 + \sqrt{1 + \frac{4}{n\,|p_2 - p_1|}}\right)^2 \tag{4.9}$$

with n denoting the sample size from Eq. (4.6) without the continuity correction, with n assumed to be the same for the two populations. [See, for example, Fleiss et al. (2003, p. 72), Ury and Fleiss (1980), or Casagrande, Pike, and Smith (1978a,b).] Note that the formula is given slightly differently in Levin and Chen (1999) since they assume that the alternative hypothesis is $p_1 > p_2$, rather than giving the form applicable for either one-sided test. (The general form for a one-sided test simply entails using $|p_2 - p_1|$, rather than using either $p_1 - p_2$ or $p_2 - p_1$.)

If unequal sample sizes are desired, as can occur due to relative costs, the desire for more precise estimates from one population, to minimize the variance of the difference of two sample proportions or their ratio (Brittain and Schlesselman, 1982), or other factors (Walter, 1977), the sample size formula with the correction factor becomes

$$n_1^* = \frac{n^*}{4}\left(1 + \sqrt{1 + \frac{2(r+1)}{nr\,|p_2 - p_1|}}\right)^2 \tag{4.10}$$

with $n_2^* = rn_1^*$ and r denoting the desired ratio of the sample sizes, n_2^*/n_1^*, with n^* given by

$$n^* = \left[\frac{z_\alpha\sqrt{(r+1)(\bar{p}\bar{q})} + z_\beta\sqrt{rp_1(1-p_1) + p_2(1-p_2)}}{r^{1/2}(p_1 - p_2)}\right]^2 \tag{4.11}$$

with $\bar{p} = (p_1 + rp_2)/(r+1)$ and $\bar{q} = 1 - \bar{p}$ (Fleiss et al. 2003, p. 76). It should be noted that Gordon and Watson (1996) questioned the utility of n_1^*, however.

If a two-sided test is to be used, the expression for n^* still applies except that now n is defined not quite as in Eq. (4.6) but is defined with z_α replaced by $z_{\alpha/2}$, as was done in Section 4.2 for the uncorrected sample size expression when a two-sided test is used.

[The reader can verify that the use of Eq. (4.9) in conjunction with $n = 296$ from Eq. (4.6) and $|p_2 - p_1| = .1$ from the current example gives $n^* = 316$.

PASS gives the same solution, as does Lenth's applet and nQuery. MINITAB, Power and Precision, and SiZ do not have a continuity correction option for testing the equality of two proportions. It can also be shown, such as by using an applet or the appropriate equation in the chapter Appendix, that the power is .8007.]

We can see from Eq. (4.8) that n^* will always be larger than n, with the percentage difference being large when $n\,|p_2 - p_1|$ is small.

Another important point is that the "remarkable accuracy" stated by Levin and Chen (1999) is relative to the sample size for Fisher's exact size. That is not entirely relevant, however, because the latter will generally not be applicable. This is because one of the assumptions of the test is that all of the marginal totals are fixed. That is, the number of subjects who are sampled from each population are fixed, as are the total number of responses for each of the two categories, such as "yes" and "no." The first of these requirements is no problem; the second one is a big problem! For this reason, papers have been written that have dispraised Fisher's exact test, including at least one that had almost that wording in the title of the paper: Berkson (1978). The point is that we would generally prefer to use an unconditional test but Fisher's exact test is a conditional test that is conditional on the four marginal totals.

D'Agostino, Chase, and Belanger (1988) considered the appropriateness of some of the commonly used tests for testing the equality of two independent binomial proportions, including Fisher's exact test. In referring to an earlier study of Upton (1982), they stated: "We confirm Upton's results that both the Fisher exact and Yates continuity correction chi-squared tests are extremely conservative and inappropriate." They demonstrated that the chi-square test *without* the continuity correction and the independent sample t-test with pooled variance have actual levels of significance that are, in most situations, close to or smaller than the nominal levels. Similarly, Gordon and Watson (1996) also advised against continuity-corrected sample size formulas.

We don't normally think about using a t-test for proportions data. It is not presented in textbooks and, indeed, D'Agostino et al. (1988) stated that "readers may be surprised by the inclusion of the t-test in an article testing binomial data."

We obtain the form of a t-test if we let a "failure" be denoted by a zero and a "success" be denoted by a one, and compute the average for each sample, which is of course also the proportion for each sample. They further stated, however, that the computation for the t-test is very close to the computation for the (uncorrected) chi-square test, which is undoubtedly due largely to the fact that the sample average is also the sample proportion.

4.2.2 Two Proportions—Fisher's Exact Test

Some software that can be used for sample size determination (including Lenth's applet and Stata) do not provide the user with the option to do an exact test for two

proportions, although this is part of the standard output of certain other software, including MINITAB. The exact test is Fisher's exact test but that test is, strictly speaking, not valid unless the four marginal totals are fixed: the sample size for population 1, the sample size from population 2, $n_1\hat{p}_1 + n_2\hat{p}_2$, and $n_1(1 - \hat{p}_1) + n_2(1 - \hat{p}_2)$. Whereas the first two will of course be fixed since the sample size from each population is being solved for, the last two will not. That is, the sum of the items from the two samples that possess a certain characteristic (e.g., lung cancer) will generally be random, not fixed.

The requirement for the fixed marginal totals results from the fact that the p-value for Fisher's exact test is obtained by computing the probability of each configuration of possible counts in the four cells of the two-way table (two samples and a "yes" or "no" categorization of each item in each sample) which is more extreme, relative to the alternative hypothesis, than what was observed. It wouldn't make any sense to do this if the marginal totals were not fixed.

Therefore, although the option for determining each sample size for the exact test is available in Power and Precision and PASS, for example, as well as other sample size determination software, the test will often be inappropriate.

If, for the sake of illustration, we assume that the test is appropriate for the example given in Section 4.2 for which $p_1 = .55$ and $p_2 = .65$, PASS shows the required common sample size to be 316, with the power being .8007.

4.2.3 What Approach Is Recommended?

D'Agostino et al. (1988) recommended the use of the t-test but also stated that "at a minimum" either the Z-test or equivalently the chi-square test should be used, without a continuity correction. Of the major sample size determination software, only PASS offers capability for a t-test. If we accept the advice of D'Agostino et al. (1988) and Gordon and Watson (1996), we will eschew approaches that employ a continuity correction and either use the sample size formula for the normal approximation in Eq. (4.6), which would indirectly also give the sample size if a chi-square test is used, or use the sample size that results from use of a t-test. For the example with p_1 assumed to be .55 and P_2 assumed to be .65, PASS gives the required sample size as $n = 296$ for use with the t-test, in agreement with the result given by PASS and certain other software for the normal approximation sample size for Eq. (4.6).

There is going to be a problem with virtually all of these approaches when population proportions are extremely small, however, such as $p = .00001$ or even less than .000001. Such proportions are encountered in low-prevalence epidemiological applications, such as those mentioned by Williams, Ebel, and Wagner (2007). They examined several sample size formulas, including those given here in Eqs. (4.9) and (4.11), and recommended a Monte Carlo approach

for determining sample size "such that the achieved power of the test closely matches the nominal value." They also pointed out that such an approach can "reduce the overall sample size by hundreds to thousands of samples while still meeting the study objectives."

4.2.4 Correlated Proportions

A paired *t*-test is used when there is a pairing of observations made on subjects, such as when two observations (measurements) are made, under different conditions, on each person, as covered in Section 3.4. Just as observations can be correlated and means can be correlated, proportions can also be correlated if proportions are computed from paired data rather than an average difference computed.

For example, the pairing could consist of people who work in a certain department for which invoice mistakes have been occurring during a 3-month period with too high of a frequency. So all of the employees are given a training program and the number of mistakes are recorded again after a second 3-month period. It is of interest to determine if the proportion of mistakes after the training program is less than the proportion before the training program. These are matched proportions because the same people are involved. (Note that testing for equality of the proportions is an alternative to determining if the average number of mistakes, for the department, after the training program is less than the average before the program.) A popular application of the test in genetics is the *transmission disequilibrium test*.

As in a test of two independent proportions, the counts or proportions could be represented in a 2×2 table. For the invoice application, the proportions would be the proportion of workers whose work was deemed satisfactory both before the program and after the program; unsatisfactory both before and after the program; satisfactory before but unsatisfactory afterward, and the reverse. If the last two are designated as b and c, respectively, the null hypothesis is that $b = c$ and one way to test this hypothesis is by using a chi-square test given by $\chi^2 = (b - c)^2/(b + c)$, which obviously measures the departure of b from c relative to the sum of the two proportions. This is *McNemar's test* (1947) using the approximation defined by the χ^2 statistic.

When PASS is used for McNemar's test, the input includes $(b - c)$ and $(b + c)$. For example, if .3 is entered for $(b - c)$ and .4 for $(b + \text{c})$, this implies that $b = .35$ and $c = .05$. With power $= .80$, and $\alpha = .05$ for a two-sided test, the required sample size is given as 36, which is the number of pairs in the sample, and the power is given as .81308. This is the exact solution, which is the default. If the approximate solution instead is requested (i.e., using the chi-square statistic), then $n = 32$. Thus, there is a noticeable difference when the two sample sizes are small, as we might expect. The PASS help file suggests using the approximate

solution initially, while stating that the approximate solution is typically about 10% less than the exact solution, as was the case with this example. The reason for this suggestion is that the algorithm for the exact solution has numerical problems when $n > 2000$. Of course, if the approximate solution is far less than 2000, then so will be the exact solution, so in this case one could have safely started with the exact approach, but not knowing that the user could have started with the approximation and then switched over to the exact approach.

PASS also has a routine for determining sample size for testing two correlated proportions in a case–control design, with this being similar to the McNemar procedure.

Literature articles in which sample size determination for testing two correlated proportions is discussed include Schork and Williams (1980). Sample sizes for designing studies with correlated binary data has also been discussed by Brooks, Cottenden, and Fader (2003). See also Nam (2011) and Royston (1993).

4.2.5 Equivalence Tests for Two Proportions

Blackwelder (1982) proposed a method for "proving" the null hypothesis, as is implied by the title of the paper. [See also Blackwelder (1998).] More specifically, the general idea was to specify a difference, D, between two population proportions that was small enough that the proportions could be declared "practically" equivalent.

The *title* of Blackwelder (1982) is a misnomer, however, because as was indicated in Section 1.1, it is never possible to literally prove any hypothesis, null or alternative, using statistical methods, whenever a sample is taken from a population. The only way to obtain "proof" is to sample the entire population, which would be impractical and cost prohibitive, and thus essentially infeasible. This should be kept in mind in reading the literature on equivalence, as equivalence testing is simply a form of hypothesis testing and does not offer the user any special powers not possessed by standard hypothesis testing.

An applet is given at `http://www.ucalgary.ca/~patten/blackwelder.html` for calculating sample sizes for equivalence testing for stated values of D and the two proportions. The formula, assuming a one-sided test, is also given there and it is

$$n = \left[\frac{z_\alpha + z_\beta}{p_1 - p_2 - D}\right]^2 [p_1(1 - p_1) + p_2(1 - p_2)] \tag{4.12}$$

To illustrate, let $p_1 = .5$, $p_2 = .6$, $D = .09$, $\alpha = .05$, and Power $= .80$. If we think of p_1 as being the proportion for a standard treatment and p_2 the expected proportion for a new treatment, if the difference turns out to be less than .09, then the two treatments will be declared equivalent. Using Eq. (4.11)

we obtain $n = 83.9$, so $n = 84$ would be used for each sample. When the applet that is at `http://www.ucalgary.ca/~patten/blackwelder.html` is used, however, $n = 83$ is obtained. The solution of $n = 84$ is what is obtained using nQuery Advisor, however. This solution can also be obtained using Power and Precision, but doing so is awkward since there is no option for solving for the sample size for power of .80, as there is with other procedures. So some trial and error would have to be used to obtain the sample size that gives the desired power. Thus, it is much easier to use nQuery Advisor for this type of sample determination problem since the solution can be obtained very easily. For this example, values of D smaller than .09 would require a larger sample size (e.g., 134 for $D = .05$), while values of D larger than .09 would lead to a smaller sample size (68 for $D = .11$). (These results are obtained using the applet and differ by one unit from the results obtained using nQuery Advisor.) The direction of the change in sample size can be seen from Eq. (4.12).

4.2.6 Noninferiority Tests for Two Proportions

Equivalence tests and noninferiority tests for a single proportion were covered in Section 4.1.3. Noninferiority tests for two proportions are often needed, such as when a new drug that is safer and/or less expensive than the standard drug is introduced and a test is to be performed that will hopefully show that the new drug is not inferior to the standard drug in terms of effectiveness. Chan (2002) considered the use of noninferiority tests for two proportions using an exact method. See also Kawasaki, Zhang, and Miyaoka (2010), who proposed new test statistics for noninferiority testing for the difference of two independent binomial proportions.

4.2.7 Need for Pilot Study?

There is only one parameter in the binomial distribution and that is the one that is being tested, unlike normality-based tests since a normal distribution has two parameters. Thus, since there is no additional parameter to be estimated, there might seem to be no need for a pilot study, such as an internal pilot study. That isn't true, however, and Friede and Kieser (2004) discussed how information from an internal pilot study could lead to a recalculation of the sample size.

4.2.8 Linear Trend in Proportions

There are various ways in which a parameter can change, with one possibility being a linear trend. This can be tested using the Cochran–Armitage (Armitage, 1955) test, which is available in PASS, with the user having the option of using

or not using a continuity correction. See also Nam (1987). For the case of two proportions, the computing formulas are the same as those given by Casagrande et al. (1978a). To illustrate, assume that there are three proportions, .25, .30, and .40, $\alpha = .05$, power $= .80$, and a one-sided test is used to detect an increasing trend. If the continuity correction is used, the sample size is 126 for each of the three groups, so the total is 378. The use of the continuity correction is recommended if the values of the covariate (such as dose values) are equally spaced. If there is unequal spacing, the PASS user can enter the spacing. If we use 1 2 4 as the spacing and the continuity correction is not used, in accordance with the recommendation of Nam (1987), the output shows a sample size of 116 for each group, for a total of 348.

4.2.9 Bayesian Method for Estimating the Difference of Two Binomial Proportions

The problem of determining sample size for estimating the difference of two binomial proportions using Bayesian methods was considered by Pham-Gia and Turkkan (2003). They briefly reviewed other published Bayesian approaches for one and two proportions and proposed a method such that either the expected length or maximum length of the highest posterior density (HPD) credible interval of the difference of the two population proportions is less than some predetermined quantity. In reviewing sample size determination methods for binomial data, Joseph et al. (1995) stated that sample sizes based on HPD intervals from the exact posterior distribution are generally smaller than sample sizes determined from the use of non-HPD intervals and/or normal approximations. In presenting Bayesian sample size methods for estimating the difference between two binomial proportions, Joseph, du Berger, and Bélisle (1997) made some important points, including the fact that Bayesian approaches generally lead to smaller sample sizes than frequentist approaches. This advantage is largely offset by the fact that, as these authors pointed out for the sample size criteria they gave in their Section 2, these are not closed-form solutions, so numerical methods must be employed to obtain the solutions. Undoubtedly, this partly explains why Bayesian methods of sample size determination are not generally available in sample size determination software.

4.3 MULTIPLE PROPORTIONS

There are various ways in which tests of the equality of more than two proportions can be performed. A common approach is to use a chi-square test and the sample size can be determined using software such as PASS and nQuery Advisor. The "effect size" must be specified with each and this is defined as the variance of the proportions divided by the product of the average proportion times one minus

the average. That is, the effect size for k proportions is given by (from nQuery Advisor)

$$\text{Effect Size} = \sum_{i=1}^{k} (p_i - \bar{p})^2/(k\bar{p}(1-\bar{p}))$$

with $\bar{p} = (\sum_{i=1}^{k} p_i)/k$.

The following example illustrates the input and shows the corresponding output.

■ EXAMPLE 4.3

Assume that a sample is to be drawn from each of three populations and a proportion computed. It is believed that the three population proportions may be approximately .40, .50, and .60, respectively, and if the proportions do differ to this extent, an experimenter wants to be able to detect these differences with power of .90, using a significance level of .05.

For nQuery Advisor, the user must compute the variance, which is part of the input, and the software then computes the effect size. The variance can easily be seen to be $.02/3 = 0.00667$ so that the effect size, as defined by this software, is $0.00667/[(.5) * (.5)] = 0.02667$, with .5 denoting the average proportion. Entering the significance level, power, number of groups (3), and average proportion (.50) leads to a sample size of $n = 159$.

The same solution is obtained using PASS. Following Cohen (1988, p. 221), PASS defines the effect size as the square root of the value of the chi-square statistic computed using the assumed proportions. That is, the chi-square statistic is

$$\sum_{i=1}^{m} \frac{(O_{ik} - E_{ik})^2}{E_{ik}}$$

with O_{ik} denoting the proportion in the ith row and kth column of a 2×3 table, and E_{ik} denoting the corresponding expected proportion. For this example all of the E_{ik} are 0.5 and the value of the chi-square statistic is $4(0.01)/0.50 = 0.08$. Therefore, the effect size that PASS uses is $\sqrt{0.08} = 0.282843$. When this value is entered for the effect size, PASS gives $n = 159$ as the solution, in agreement with the solution given by nQuery Advisor, for example. ■

There is often a need to test the equality of multiple proportions, such as when comparing the work efficiency of several workers in a department. Ryan (2011, p. 592) gave an example, drawn from the author's consulting experience, involving the comparison of 20 (human) harvesters, with a desire to identify any

who stand out in a positive way and whose performance traits should perhaps be studied by other workers, as well as those workers who are significantly below average and thus might need to be retrained or reassigned. This is best performed with a graphical test, whereas the chi-square test is obviously not a graphical test. Ryan (2011) showed that one of the human harvesters was significantly better than the average performance of all of them, and one harvester was significantly worse than the average performance.

An ANOM plot (but not for the data just mentioned), produced by MINITAB, showing one point above the upper decision line (UDL) is shown in Figure 4.3.

Another example of ANOM applied to proportions data was given by Homa (2007), in which she compared the referral rates of 17 providers, with spine patients who scored low on a mental health test being referred to Behavioral Medicine Services. This is not a sample size determination problem, though, since these are observational data and the sample size for each provider is the number of patient visits, which of course can't be set or predetermined. Nevertheless, it is of interest to look at the sample sizes, which of course varied among providers relative to the objective of the study, which was "to demonstrate through a case study how an analysis of means chart can be used to compare groups and to advocate the usefulness of this method in improvement work" (Homa, 2007). The ANOM analysis showed that three providers had a significantly higher rate than the overall average (using a .05 upper decision line), and three providers had a significantly lower rate (with a .05 lower decision line). Actually, multiple

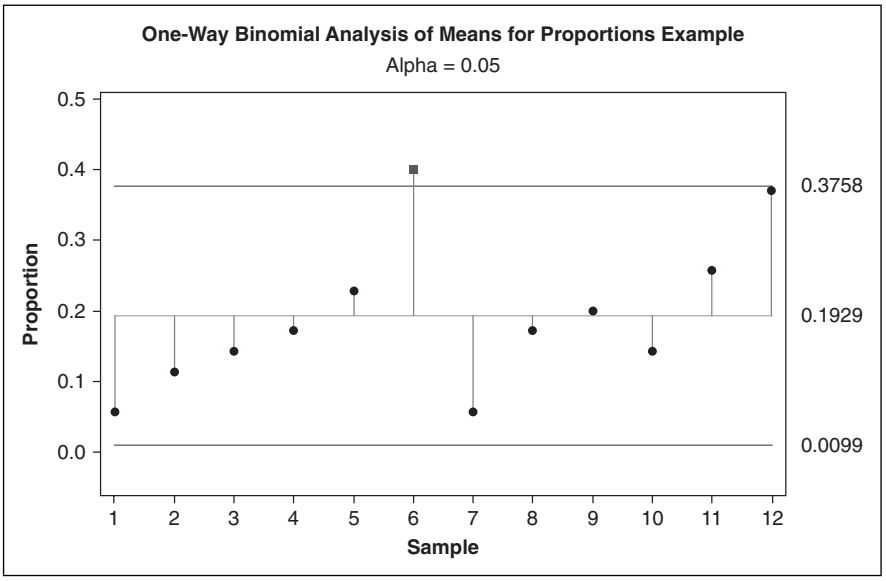

Figure 4.3 ANOM display for testing the equivalence of 12 proportions.

ANOM charts were constructed, being defined on each of several patient characteristics. The analysis led to some questions that would have to be addressed with further study.

There is almost certainly no software available for determining sample size for ANOM, and there is very little in the literature on the subject. Consequently, it is necessary to (attempt to) derive the sample size expression and we can proceed as follows. We first need to determine what magnitude of a difference we wish to detect. Since the relative strength of ANOM for proportions is in detecting an extreme proportion, the sample size might be determined for detecting a single proportion that might be a specified multiple of $\bar{p}$, such as 1.5 or 0.5, with $\bar{p}$ denoting, as before, the average of all of the plotted proportions and the other $(k - 1)$ proportions assumed to be representable by $\bar{p}$.

We immediately encounter somewhat of a potential mismatch, however, because the value of α is for testing *all* of the plotted proportions against the midline represented by $\bar{p}$, whereas we might be inclined to define β relative to testing a single large or small proportion, ignoring the other proportions.

As indicated earlier in this section, an alternative to using ANOM for testing the equality of proportions is to use a chi-square test, for which sample size determination software *is* available and was illustrated. A chi-square test is not as sensitive as ANOM for detecting a few extreme deviations, however, so if we are interested in extreme proportions in either direction, as in the human harvesters example, ANOM would be a better choice. Another problem is that it may be more difficult to specify a *logical* "effect" size for which it is desirable to detect, since a chi-square test produces a single number, as opposed to the ANOM display which shows each proportion plotted relative to the decision lines and the center line. Cohen (1988) recommended specifying an effect of .10, with the effect defined simply as χ^2/N, as discussed previously, with N denoting the sum of the counts in the $k \times 2$ table, with k denoting the number of proportions that are plotted. This effect size definition does not have any obvious intuitive meaning, however, so there are various reasons why ANOM for proportions should be used and why software should be available for determining sample size for such a procedure.

4.4 MULTINOMIAL PROBABILITIES AND DISTRIBUTIONS

The multinomial probability distribution is an extension of the binomial distribution. Although there has been some research on sample size determination for multinomial probabilities, notably by Adcock (1987, 1992, 1993) who considered Bayesian approaches and by Guenther (1977a) who considered sample size and power for hypotheses concerning multinomial probabilities with chi-square tests, sample size determination for estimating multinomial probabilities won't be pursued here. Instead, references are given for interested readers. Additional

references include Bromaghin (1993) and Lee, Song, Kang, and Ahn (2002). See Nisen and Schwertman (2008) for sample size determination for chi-square tests used to test the equality of multinomial distributions.

4.5 ONE RATE

We will first consider a Poisson distribution, given by

$$f^*(x) = \frac{\exp(-\lambda)\lambda^x}{x!}$$

with λ denoting the Poisson parameter (rate) and of course X is the random variable with realization denoted by x. Although $\sum_{i=1}^{n} X_i$ has a Poisson distribution with parameter $n\lambda$, $\sum_{i=1}^{n} X_i/n$ does not. Therefore, one approach would be to use $\sum_{i=1}^{n} X_i$ as the test statistic, and Lenth's applet does use this test statistic. Assume that we wish to perform a one-sided test, with $\alpha = .05$, and want the power to be .90 for detecting an increase in λ from 1 to 1.5. Let λ_0 denote the hypothesized value of 1 and $\lambda_1 = 1.5$ denote the value that is to be detected.

We want to determine n such that $P(\sum_{i=1}^{n} X_i > c \mid \lambda = n\lambda_1) = .90$, with c determined such that $P(\sum_{i=1}^{n} X_i > c \mid \lambda = n\lambda_0) = .05$ (i.e., c is the critical value of the test). Since there is no expression for c that is a function of n, there is no equation to be solved for n. Furthermore, since the random variable is discrete, we generally cannot hit $\alpha = .05$ exactly, so we will use $\alpha < .05$. Since there is no equation to solve for n, as stated, it is necessary to use software and not all sample size determination software has this capability.

PASS does have the capability (under "means" in its main menu) but it does not perform an exact test since it determines sample size following Guenther (1977b), which utilizes the relationship between the chi-square and Poisson distributions [e.g., see Ulm (1990)]. Specifically, PASS determines the sample size by using the interval

$$\frac{\chi^2_{2d,1-\beta}}{2\lambda_1} \le n \le \frac{\chi^2_{2d,\alpha}}{2\lambda_0} \qquad d = 1, 2, 3, \ldots$$

with λ_0 denoting the hypothesized value of λ and λ_1 denoting the value of λ that is to be detected, with the first subscript on χ^2 denoting the degrees of freedom and the second subscript denoting the percentile of the chi-square distribution with $2d$ degrees of freedom. The value of n is found by increasing the value of d until the left interval endpoint is less than the right endpoint and the interval contains at least one integer. To illustrate, with the current example having $\lambda_0 = 1$ and $\lambda_1 = 1.5$, for a one-sided test with $\alpha = .05$ and power $= .90$, PASS gives

$n = 44$. This solution can be seen by observing that at $d = 56$ so that $2d = 112$, $\chi^2_{112,.90} = 131.558$ and $\chi^2_{112,.05}| = 88.5704$, so that the left endpoint is 43.8527 and the right endpoint is 44.2852, so that the interval barely contains $n = 44$. (It can be shown that the inequality is not satisfied at $2d = 110$, thus verifying that the solution is that obtained when $2d = 112$, so that $n = 44$.)

It should also be noted that a two-sided test is not a menu option in PASS, but the sample size for a two-sided test can be determined by specifying $\alpha/2$, as pointed out by the PASS help system.

Lenth's applet also has sample size capability for Poisson rates, as does Release 16 of MINITAB (but not in prior releases). Using Lenth's applet also results in $n = 44$ as the solution. The value of c is 55 as it can be shown that $P(\sum_{i=1}^{n} X_i > 55 \mid \lambda = (44)(1)) = .0456$, and $P(\sum_{i=1}^{n} X_i > 55 \mid \lambda = (44)(1.5)) = .9046$. Lenth gives the critical value as $c = 54$, however, whereas 55 is the largest value that would result in the null hypothesis not being rejected.

A normal approximation approach could be used, analogous to what is done with proportions, but that approximation will not work well when $\lambda < 5$, as it was in this example. An exact test is thus the preferred approach, with either PASS or Lenth's applet being good options as neither nQuery Advisor nor Power and Precision has the capability of either the exact approach or a normal approximation approach, nor does G*Power, Stata, or SiZ.

By comparison, the MINITAB solution for the preceding example is $n = 42$. For that solution, MINITAB gives the actual power of .9037 but there is no value of c that will produce .9037 for $\lambda = 42(1.5) = 63$. Thus, MINITAB could not have used an exact test and thus it is very likely the solution is obtained using a normal approximation approach, as the user does not have an option for the choice of test. It will often be desirable to first use a square root transformation for Poisson data before applying a normal approximation. Mathews (2010) gives the sample expression when that is done. Using that expression produces $n = 42.39$ so that $n = 43$ would be used. Thus, it is very probable that MINITAB used a normal approximation without a square root transformation.

Regarding other approaches, van Belle (2008, p. 420) gave a rule of thumb for the sample size that is the application of Eq. (3.7) to the Poisson case and also assuming the use of the square root transformation of Poisson data since $\sqrt{X}$ is approximately $N(\sqrt{\lambda}, 0.25)$. Substitution of these quantities into Eq. (3.7) produces $n = 4/(\sqrt{\lambda_0} - \sqrt{\lambda_1})^2$. Note that this could be used only for a power of .80 and a significance level for a two-sided test of .05. For detecting an increase in λ from 9 to 16, this formula gives $n = 16$, but this applies to the transformed variable $\sqrt{X}$, not to X. Therefore, it cannot be compared against the results obtained using sample size determination software for the Poisson parameter, as such calculations would be for the Poisson distribution, for which the mean and the variance are assumed to be equal. [It can be shown, however, that the sample size approximation is poor for a change in a normal mean from $\sqrt{9} = 3$ to $\sqrt{16} = 4$ when

the variance is assumed to be 0.25, as Eq. (3.2) or the use of software produces $n = 2$. Recall that Eq. (3.7) has been criticized as being inaccurate, however, as discussed in Section 3.6. Thus, it is probably best not to use approximations based on that formula. Furthermore, of course, two approximations are being used in applying the van Belle rule of thumb, as an adequate approximation to normality is also being assumed. With today's computing power it is both unnecessary and undesirable to rely heavily on approximations.]

4.5.1 Pilot Study Needed?

As with the binomial distribution, the Poisson distribution has only one parameter, so there is no nuisance parameter to be estimated. Therefore, assuming that there is a good handle on what the null hypothesis should be, a pilot study might be considered to determine if the rate specified in the alternative hypothesis is reasonable, although what rate the user wants to detect might already have been determined with some certainty. Thus, there will not necessarily be a strong motivation for a pilot study when testing a Poisson rate, except in clinical trials work when there are three phases. Friede and Schmidli (2010) presented formulas for adjusting the sample size for Poisson-distributed data and for Poisson-distributed data with overdispersion (excess variability). In a thesis on sample size determination with an auditing application, Pedersen (2010) used a zero-inflated Poisson model (a model used to overcome problems posed by an excessive number of zeros with Poisson data) and considered sample sizes for both frequentist and Bayesian approaches.

4.6 TWO RATES

When a company has two (or more) plants, and one plant seems to be having trouble making a particular product, it would be desirable to compare the rate of nonconformities for the plants and do this repeatedly at specified intervals and to do this until the null hypothesis of equal rates cannot be rejected.

Specifically, assume that the rate at the better of two plants is 2 for some time period and the objective is to determine the (common) sample size so that a rate of 4 at the other plant is to be detected with probability (power) of .90. This might be addressed as a one-sample problem, using 2 as the desired value since that is the value for the first plant and the company personnel would then determine what would constitute an unacceptable rate for the second plant relative to that baseline value. This would permit the use of software such as MINITAB or Lenth's applet, as illustrated in Section 4.4. Unfortunately, there is apparently very little commercial software that can be used to determine sample size for testing the equality of two Poisson rates.

MINITAB does have a two-sample Poisson rate procedure in Release 16, which uses an unspecified iterative algorithm for determining the sample size. For example, if we specify a null hypothesis in which two Poisson rates are equal, then the "baseline rate" used by MINITAB, which is the ratio of the two Poisson rates under the null hypothesis, is 1.0. If the first Poisson rate is 0.8 and the second Poisson rate is 0.6 under the alternative hypothesis, so that the comparison rate is 0.8/0.6 = 1.33, with $\alpha = .05$ and power = .90, the output gives a sample size of $n = 180$ for each of the two samples for a one-sided test.

This agrees with the formula for the sample size of the unconstrained maximum likelihood test given by Gu, Ng, Tang, and Schucany (2008), which is

$$n = \left[\frac{\left(c/\rho + c^2\right)\left(z_\alpha + z_\beta\right)^2}{(1-c)^2} \right]$$

with $\rho = 1$ when the null hypothesis is that the sampling rates are equal and the two fixed sampling frames are also equal, and c is the ratio of the two Poisson rates under the null hypothesis divided by the ratio of the first rate to the second rate under the alternative hypothesis, provided that the ratio exceeds the ratio under the null hypothesis. Thus, for this example,

$$\begin{aligned} n &= \left[\frac{(0.75 + 0.75)^2 (1.645 + 1.28155)^2}{(1 - .75)^2} \right] \\ &= 179.859 \end{aligned}$$

so that $n = 180$ would be used.

PASS also has the capability for a two-sample Poisson test, with its test being one of the four types of tests covered by Gu et al. (2008). The user has five options for the test statistic, with the unconstrained maximum likelihood test in which the two Poisson means are estimated separately being the default. Unlike MINITAB's routine, the two-sample Poisson routine in PASS requires that the ratio of the two means be specified in both the null and alternative hypotheses. The use of PASS for the unconstrained maximum likelihood test also gives $n_1 = n_2 = 300$, in agreement with MINITAB. (The same sample size is obtained if use of the constrained maximum likelihood test is assumed, for which the two Poisson means are estimated jointly.)

Gu et al. (2008), however, performed simulations and found that the asymptotic test derived from the variance-stabilizing transformation proposed by Huffman (1984) was the most reliable asymptotic test, being conservative but having high power. This is the last of the five options available in PASS.

Gu et al. (2008) gave the sample size formulas for all of the methods that they covered, with the formula for their preferred test, based on the variance-stabilizing transformation, being

$$n = \left[\frac{z_\alpha \sqrt{c/\rho + c} + z_\beta \sqrt{1 + c/\rho}}{2\left(1 - \sqrt{c}\right)} \right]^2 - \frac{3}{8}$$

with $\rho = R/d$ and R is the ratio under the null hypothesis of the first Poisson mean (i.e., rate) divided by the second Poisson mean, $c = R/R'$, with R' denoting the ratio of the Poisson means under the alternative hypothesis, and $d = t_1/t_0$, with t_1 and t_0 denoting the fixed sampling frames (i.e., time intervals) for the two Poisson processes.

Continuing with the example given earlier in this section, for this test PASS gives the sample size as 319 for each sample. Except for this result, the sample sizes are almost identical (300, 300, 302, and 296, respectively) for the first four options in PASS.

Using hand computation, we obtain

$$\begin{aligned} n &= \left[\frac{z_\alpha \sqrt{c/\rho + c} + z_\beta \sqrt{1 + c/\rho}}{2\left(1 - \sqrt{c}\right)} \right]^2 - \frac{3}{8} \\ &= \left[\frac{1.645\sqrt{0.75 + 0.75} + 1.28155\sqrt{1 + .75}}{2\left(1 - \sqrt{.75}\right)} \right]^2 - \frac{3}{8} \\ &= 191.338 \end{aligned}$$

so $n = 192$ would be used. Notice that this sample size is almost 7% greater than the sample size obtained for the unconstrained maximum likelihood test.

Nelson (1991) gave a normal approximation approach and provided a computer program (written in BASIC) for performing the computations, but a normal approximation approach won't work when the Poisson rates are small, as will be the case in many applications.

A method for determining sample size and power in comparing two Poisson rates has been given by Thode (1997) and by Hand, Stamey, and Young (2011). See also Stamey and Katsis (2007) for the case of rates that are underreported. Gu et al. (2008) examined the proposed methods for testing two Poisson rates when the null hypothesis for the ratio of the rates does not equal 1. They provided sample size formulas for each approach and gave their recommendations. Shiue and Bain (1982) considered the determination of the experiment length needed to achieve a specified power for an approximate test when the two intervals are equal and when they are unequal. Ng and Tang (2005) considered methods for

testing the equality of Poisson means when the intervals are unequal and gave sample size formulas for that case.

Although not well known, the distribution of the difference of two Poisson random variables is due to Skellam (1946), and is referred to as the Skellam distribution. Menon, Massaro, Lewis, Pencina, Wang, and Lavin (2011) used this exact distribution in a proposed procedure. They gave an example to illustrate their methodology, which is as follows. Suppose there is interest in the relapse rate per year of an infectious disease for two treatments. The rate for the standard treatment is expected to be 6.25, and 5.15 for the new treatment. Assume that the target power for detecting this difference is .90 and a one-sided test with $\alpha = .05$ is to be performed. They gave $n = 87$ as the sample size to be used for each group. (Their method is in the `skellam` package that is in *R*.) By comparison, when MINITAB is used, the solution is $n = 81$, so the exact procedure requires a slightly larger sample size.

4.7 BAYESIAN SAMPLE SIZE DETERMINATION METHODS FOR RATES

Bayesian sample size determination methods for both one and two Poisson rates were given by Stamey, Young, and Bratcher (2006), who gave two actual quality control examples and compared the results to sample sizes determined using frequentist methods.

4.8 SOFTWARE

Unlike the case with sample means, there is some disagreement among software that handles sample size determination for proportions, as was indicated in Section 4.1.2. As another example, for the case of a hypothesized proportion of .50, an upper-tailed test with no continuity correction, and an assumed true proportion of .60, with power of .80, Lenth's applet gives a sample size of 153, as does MINITAB, Stata, and Power and Precision, but PASS gives 146. There is disagreement because the latter uses a combination of a normal approximation approach and an exact approach, whereas the others use a normal approximation approach throughout.

There are, of course, applets that can be used for determining sample size with one or two proportions, in addition to Lenth's applet, although not all of these can be recommended. One of these other applets is at `http://statpages.org/proppowr.html` but it is not clear what alternative hypothesis is being assumed.

There is very little available software for determining sample size for testing the equality of two Poisson rates. Lenth's applet can be used for one rate, but not for the test of the equality of two rates. PASS and MINITAB can be used for

this purpose, however, as was illustrated, but sample size for two rates cannot be determined using either nQuery Advisor, Power and Precision, G*Power, or Stata.

4.9 SUMMARY

This chapter contained considerable discussion of the options that a user has in determining sample size for either one or two proportions. Certain software do not provide options, however, and the user needs to be cognizant of the method that is being used to determine sample size. Software for rates is not as plentiful as software for proportions, and this is especially true for two rates.

APPENDIX

(a) Derivation of Eq. (4.2)—No Continuity Correction

As with any sample size formula derivation, the first step is to obtain the expression for β since the formula will be a function of β. The test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} \tag{A.1}$$

which is approximately $N(0,1)$ when n is large only if $p = p_0$, with the latter denoting the hypothesized value of p. The term "power" has meaning only if $p \neq p_0$ and as pointed out in Chapter 1, null hypotheses are almost always false anyway. We will let p_1 denote the actual value of p.

The sample size expression in Eq. (4.2) does not depend on the direction of the inequality, but for the derivation we will assume that it is an upper-tailed test so that the critical value of Z is z_α. The null hypothesis is erroneously not rejected if

$$\frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} < z_\alpha \tag{A.2}$$

Since the left side of Expresssion (A.2) is not approximately $N(0,1)$ when $p = p_1$, it is necessary to modify that expression to create an expression that is approximately $N(0,1)$. Since $E(\hat{p}) = p_1$, the expected value of the left side of expression

(A.2) is $(p_1 - p_0)/\sqrt{p_0(1-p_0)/n}$. Since $\hat{p}$ is the only random variable on the left side, the variance of the expression *might seem to be*

$$\text{Var}\left(\frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}\right) = \frac{p_1(1-p_1)/n}{p_0(1-p_0)/n}$$

Therefore, the appropriate standardization would be

$$\frac{\dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} - \dfrac{p_1 - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}}{\sqrt{\dfrac{p_1(1-p_1)/n}{p_0(1-p_0)/n}}} < \frac{z_\alpha - \dfrac{p_1 - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}}{\sqrt{\dfrac{p_1(1-p_1)/n}{p_0(1-p_0)/n}}}$$

$$< \frac{z_\alpha\sqrt{\dfrac{p_0(1-p_0)}{n}} - (p_1 - p_0)}{\sqrt{\dfrac{p_1(1-p_1)}{n}}} = C(\text{say})$$

Thus, $\beta = \Phi(C)$, so that $C = \Phi^{-1}(\beta)$. Therefore,

$$z_\alpha\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} - \frac{p_1 - p_0}{\sqrt{\dfrac{p_1(1-p_1)}{n}}} = \Phi^{-1}(\beta) = z_\beta = -z_\beta$$

so

$$-\frac{p_1 - p_0}{\sqrt{\dfrac{p_1(1-p_1)}{n}}} = -z_\beta - z_\alpha\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}$$

Solving this last equation for n produces

$$n = \left[\frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_1(1-p_1)}}{p_1 - p_0}\right]^2 \tag{A.3}$$

For a two-sided test, z_α would be replaced by $z_{\alpha/2}$.

The sample size expression given by Eq. (A.3) is what is given by Fleiss et al. (2003) and used by the software nQuery as well as Russ Lenth's software. These book authors, Lenth, and the developers of the single proportion routine in nQuery would undoubtedly agree with the derivation given here. However, under the classical Neyman–Pearson theory of hypothesis testing, parameters that appear in test statistics, and hence in sample size formulas that are derived from those test statistics, are set equal to their values under the null hypothesis. When this is done here, Eq. (A.3) becomes

$$n = \left[\frac{(z_\alpha + z_\beta)\sqrt{p_0(1-p_0)}}{p' - p_0}\right]^2 \tag{A.4}$$

which is the expression given by Chow et al. (2008, p. 86). Thus, there is disagreement about the formula for *n,* although the sample sizes should not differ very much between the two formulas unless p' differs more than slightly from p_0. Fleiss et al. (2003) derived their sample size expression from the expression for power at p'. Since this is based on the assumption that p' is the true proportion, this results in p' being in more than just the denominator in the expression for n.

(b) Derivation of Eq. (4.6)—No Continuity Correction

We can proceed analogously to the derivation of Eq. (4.1). That is, we start with the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\dfrac{2\bar{p}(1-\bar{p})}{n}}} \tag{A.5}$$

with 0 representing the value of $p_1 - p_2$ under the null hypothesis, n is the assumed common sample size for samples from the two populations, and $\bar{p} = (\hat{p}_1 + \hat{p}_2)/2$ for a common sample size.

As in the derivation of the sample size expression for a single proportion, we will assume an upper-tailed test so that the null hypothesis of $p_1 = p_2$ should be rejected in favor of the alternative hypothesis, $p_1 > p_2$. Under the alternative hypothesis, the expected value and the variance of the Z-statistic in Eq. (A.3) are approximately

$$\frac{p_1 - p_2}{\sqrt{\dfrac{2\bar{p}^*(1-\bar{p}^*)}{n}}} \quad \text{and} \quad \sqrt{\frac{[p_1(1-p_1) + p_2(1-p_2)]/n}{2\bar{p}^*(1-\bar{p}^*)/n}}$$

respectively, with $\bar{p}^* = (p_1 + p_2)/2$. The results are approximate because the denominator in Eq. (A.3), which contains the random variable $\bar{p}$, is treated as a constant. Then,

$$\frac{z_\alpha - \dfrac{p_1 - p_2}{\sqrt{\dfrac{2\bar{p}^*(1-\bar{p}^*)}{n}}}}{\sqrt{\dfrac{[p_1(1-p_1) + p_2(1-p_2)]/n}{2\bar{p}^*(1-\bar{p}^*)/n}}} = \Phi^{-1}(\beta)$$

so

$$\frac{\sqrt{\dfrac{2\bar{p}^*(1-\bar{p}^*)}{n}}z_\alpha - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1) + p_2(1-p_2)}{n}}} = \Phi^{-1}(\beta) = z_\beta = -z_\beta$$

Thus,

$$z_\alpha \left[\frac{\sqrt{2\bar{p}^*(1-\bar{p}^*)}}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}\right] = -z_\beta + \sqrt{n}\left[\frac{(p_1 - p_2)}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}\right]$$

and solving this equation for n produces Eq. (4.6), after substituting $(p_1 + p_2)/2$ for $\bar{p}^*$ That is,

$$n = \left(\frac{z_\alpha\sqrt{2\bar{p}^*(1-\bar{p}^*)} + z_\beta\sqrt{p_1(1-p_1) + p_2(1-p_2)}}{(p_1 - p_2)}\right)^2$$

before $(p_1 + p_2)/2$ is substituted for $\bar{p}^*$, with that substitution then producing Eq. (4.6).

(c) Explanation of Eq. (4.9)—Using Continuity Correction for Two Proportions

In using the continuity correction for the two-proportion case, we proceed analogously to the one-sample case. Specifically, if the alternative hypothesis is $p_1 - p_2 > 0$, then we should think of p_1 "starting" (loosely speaking, in terms of continuity correction) at $p_1 - 1/2n$. Similarly, we should think of p_2 as "ending" at $p_2 + 1/2n$. The numerator of the statistic should then be $p_1 - 1/2n - (p_2 + 1/2n) = p_1 - p_2 - 1/n$. If the alternative hypothesis had been $p_1 - p_2 < 0$, the numerator would have been $p_1 - p_2 + 1/n$. Let $p_1 - p_2 = a$ for the case $p_1 - p_2 > 0$ and $-a$ for the case $p_1 - p_2 < 0$. We then have $a - 1/n$ and $-a + 1/n$, so that the

negative of the latter is the former. The sample size will, of course, be the same for the two cases, so we want to use $|-a+1/n|$ in computing the required sample size. As explained by Fleiss et al. (2003, p. 72), the sample size formula gives the impression that only one continuity correction is being employed because the second continuity correction effectively cancels out the round-off error. [See also the explanation of this given by Levin and Chen (1999, p. 64).]

REFERENCES

Adcock, C. J. (1987). A Bayesian approach to calculating sample sizes for multinomial sampling. *The Statistician*, **36**, 155–159.

Adcock, C. J. (1992). Bayesian approaches to the determination of sample sizes for binomial and multinomial sampling—some comments on the paper by Pham-Gia and Turkkan. *The Statistician*, **41**, 399–401.

Adcock, C. J. (1993). A improved Bayesian approach for calculating sample sizes for multinomial sampling. *The Statistician*, **42**(2), 91–95.

Afshartous, D. (2008). Sample size determination for binomial proportion confidence intervals: An alternative perspective motivated by a legal case. *The American Statistician*, **62**, 27–31.

Agresti, A. and B. A. Coull (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52**(2), 119–126.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**(3), 375–386.

Berkson, J. (1978). In dispraise of the exact test. *Journal of Statistical Planning and Inference*, **2**, 27–42.

Blackwelder, W. (1982). Proving the null hypothesis in clinical trials. *Controlled Clinical Trials*, **3**, 345–353.

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

Brittain, E. and J. J. Schlesselman (1982). Optimal allocation for the comparison of proportions. *Biometrics*, **38**, 1003–1009.

Bromaghin, J. F. (1993). Sample size determination for interval estimation of multinomial probabilities. *The American Statistician*, **47**(3), 203– 206.

Brooks, R. J., A. M. Cottenden, and M. J. Fader (2003). Sample sizes for designed studies with correlated binary data. *Journal of the Royal Statistical Society, Series D*, **52**, 539–551.

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.

Casagrande, J. T., M. C. Pike, and P. G. Smith (1978a). An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*, **34**, 483–486.

Casagrande, J. T., M. C. Pike, and P. G. Smith (1978b). The power function of the "exact" test for comparing two binomial proportions. *Applied Statistics*, **78**, 176–180.

Cesana, B. M., G. Reina, and E. Marubini (2001). Sample size for testing a proportion in clinical trials: A "two-step" procedure combining power and confidence interval expected width. *The American Statistician*, **55**, 265–270.

Chan, I. S. F. (2002). Power and sample size determination for noninferiority trials using an exact method. *Journal of Biopharmaceutical Statistics*, **12**, 457–469.

Chernick, M. R. and C. Y. Liu (2002). The saw-toothed behavior of power versus sample and software solutions: Single binomial proportion using exact methods. *The American Statistician*, **56**(2), 149–155.

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

Clopper, C. J. and E. S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Philadelphia: Lawrence Erlbaum Associates.

Conover, W. J. (1980). *Practical Nonparametric Statistics*, 2nd edition. New York: Wiley.

D'Agostino, R. B., W. Chase, and A. Belanger (1988). The appropriateness of some common procedures for testing equality of two independent binomial proportions. *The American Statistician*, **42**(3), 198–202.

Farrington, C. P. and G. Manning (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, **9**, 1447–1454.

Fleiss, J. L., A. Tytun, and H. K. Ury (1980). A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, **36**, 343–346.

Fleiss, J. L., B. Levin, and M. C. Paik (2003). *Statistical Methods for Rates and Proportions*, 3rd edition. Hoboken, NJ: Wiley.

Friede, T. and M. Kieser (2004). Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics*, **3**(4), 269–279.

Friede, T. and H. Schmidli (2010). Blinded sample size reestimation with count data: Methods and applications in multiple sclerosis. *Statistics in Medicine*, **29**, 1145–1156.

Gould, A. L. (1983). Sample sizes required for binomial trials when the true response rates are estimated. *Journal of Statistical Planning and Inference*, **8**, 51–58.

Gordon, I. and R. Watson (1996). The myth of continuity-corrected sample size formulae. *Biometrics*, **52**, 71–76.

Gu, K., H. K. T. Ng, M. L. Tang, and W. Schucany (2008). Testing the ratio of two Poisson rates. *Biometrical Journal*, **50**(2), 283–298.

Guenther, W. C. (1977a). Power and sample size for approximate chi-square tests. *The American Statistician*, **31**, 83–85.

Guenther, W. C. (1977b). *Sampling Inspection in Statistical Quality Control*. Griffin's Statistical Monographs, No. 37. London: Griffin.

Hand, A. L., J. D. Stamey, and D. M. Young (2011). Bayesian sample-size determination for two independent Poisson rates. *Computer Methods and Programs in Biomedicine*, **104**(2), 271–277.

Homa, K. (2007). Analysis of Means used to compare providers' referral patterns. *Quality Management in Health Care*, **16**(3), 256–264.

Huffman, M. (1984). An improved approximate two-sample Poisson test. *Applied Statistics*, **33**(2), 224–226.

Joseph, L., D. B. Wolfson, and R. du Berger (1995). Some comments on Bayesian sample size determination. *The Statistician*, **44**(2), 167–171.

Joseph, L., R. Du Berger, and P. Bélisle (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, **16**(2), 769–781.

Kawasaki, Y., F. Zhang, and E. Miyaoka (2010). Comparisons of test statistics for noninferiority test for the difference between two independent binomial proportions. *American Journal of Biostatistics*, **1**(1), 23–30.

Khuri, A. I. (1978). A conservative sample size for the comparison of several proportions. *Communications in Statistics—Theory and Methods*, **7**(13), 1283–1293.

Lee, M.-K., H.-H. Song, S.-H. Kang, and C. W. Ahn (2002). The determination of sample sizes in the comparison of two multinomial proportions from ordered categories. *Biometrical Journal*, **44**, 395–409.

Levin, B. and X. Chen (1999). Is the one-half continuity correction used once or twice to derive a well-known approximate sample size formula to compare two independent binomial distributions? *The American Statistician*, **53**, 62–66.

Levy, P. S. and S. Lemeshow (1991). *Sampling of Populations: Methods and Applications*. New York: Wiley.

Mathews, P. (2010). *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Fairport Harbor, OH: Mathews Malnar and Bailey, Inc.

McNemar, Q. (1947). Note on the sampling error of the difference of two correlated proportions or percentages. *Psychometrika*, **12**(2), 153–157.

Menon, S., J. Massaro, J. Lewis, M. Pencina, Y.-C. Wang, and P. Lavin (2011). Sample size calculation for Poisson endpoint using the exact distribution of difference between two Poisson random variables. *Statistics in Biopharmaceutical Research*, **3**(3), 497–504.

M'Lan, C. E., L. Joseph, and D. B. Wolfson (2008). Bayesian sample size determination for binomial proportions. *Bayesian Analysis*, **3**(2), 269–296.

Nam, J.-M. (1987). A simple approximation for calculating sample sizes for detecting linear trends in proportions. *Biometrics*, **43**, 701–705.

Nam, J.-M. (2011). Power and sample size requirements for non-inferiority in studies comparing two matched proportions where the events are correlated. *Computational Statistics and Data Analysis*, **55**, 2880–2887.

Nelson, L. S. (1991). Power in comparing Poisson means: II. Two-sample test. *Journal of Quality Technology*, **23**(2), 163–166.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, **17**, 857–872.

Ng, H. K. T. and M. L. Tang (2005). Testing the equality of two Poisson means using the rate ratio. *Statistics in Medicine*, **24**, 955–965.

Nisen, J. A. and N. C. Schwertman (2008). A simple method of computing the sample size for chi-square test for the equality of multinomial distributions. *Computational Statistics and Data Analysis*, **52**(11), 4903–4908.

Pedersen, K. (2010). Sample size determination in auditing accounts receivable using a zero-inflated Poisson model. M.S. thesis, Worcester Polytechnic Institute.

Pham-Gia, T. and N. Turkkan (1992). Sample size determination in Bayesian analysis. *The Statistician*, **41**, 389–397.

Pham-Gia, T. and N. Turkkan (2003). Determination of exact sample sizes in the Bayesian estimation of the difference of two proportions. *The Statistician*, **52**, 131–150.

Royston, P. (1993). Exact conditional and unconditional sample size for pair-matched studies with binary outcome: A practical guide. *Statistics in Medicine*, **12**, 699–712.

Ryan, T. P. (2007). *Modern Engineering Statistics.* Hoboken, NJ: Wiley.

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

Samuels, M. L. and T. C. Lu (1992). Sample size requirements for the back-of-the-envelope binomial confidence interval. *The American Statistician*, **46**(3), 228–231.

Schader, M. and F. Schmid (1989). Two rules of thumb for the approximation of the binomial distribution by the normal distribution. *The American Statistician*, **43**, 23–24.

Schork, M. and G. Williams (1980). Number of observations required for the comparison of two correlated proportions. *Communications in Statistics—Simulation and Computation*, **B9**(4), 349–357.

Shiue, W.-K. and L. J. Bain (1982). Experiment size and power comparisons for two-sample Poisson tests. *Applied Statistics*, **31**,130–134.

Skellam, J. G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society, Series A*, **109**, 296.

Stamey, J. and A. Katsis (2007). Sample size determination for comparing two Poisson rates with underreported counts. *Communications in Statistics—Simulation and Computation*, **36**(3), 483–492.

Stamey, J., D. M. Young, and T. L. Bratcher (2006). Bayesian sample-size determination for one and two Poisson rate parameters with applications to quality control. *Journal of Applied Statistics*, **33**(6), 583–594.

Suissa, S. and J. J. Shuster (1985). Exact unconditional sample sizes for the 2×2 binomial trial. *Journal of the Royal Statistical Society, Series A*, **148**, 317–327.

Thode, H. C. Jr. (1997). Power and sample size requirements for tests of differences between two Poisson rates. *The Statistician*, **46**(2), 227–230.

Tobi, H.; P. B. van den Berg, and L. T. de Jong-van den Berg (2005). Small proportions: What to report for confidence intervals? *Pharmacoepidemiology Drug Safety*, **14**(4), 239–247. Erratum: **15**(3), 211.

Ulm, K. (1990). A simple method to calculate the confidence interval of a standardized mortality ratio. *American Journal of Epidemiology*, **131**(2), 373–375.

Upton, G. J. G. (1982). A comparison of alternative tests for the 2×2 comparative trial. *Journal of the Royal Statistical Society, Ser A*, **145**, 86–105.

Ury, H. K. and J. L. Fleiss (1980). On approximate sample sizes for comparing two independent proportions with the use of Yates' correction. *Biometrics* **36**, 347–351.

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

Vollset, S. E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, **12**(9), 873–890 (author's reply: **13**, 809–824).

Walter, S. D. (1977). Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *American Journal of Epidemiology*, **105**, 387–397.

Williams, M. S., E. D. Ebel, and B. A. Wagner (2007). Monte Carlo approaches for determining power and sample size in low-prevalence applications. *Preventive Veterinary Medicine*, **82**, 151–158.

EXERCISES

4.1. Derive the sample size of 296 that was given, just after Eq. (4.6), in an example in Section 4.2.

4.2. The website for the American Academy of Periodontology states that 30 percent of adults over the age of 50 have periodontal disease. Literature handed out by a dentist (true story) stated that 3 out of 4 adults over the age of 35 have periodontal disease. The state dental board doesn't believe the literature but decides to perform a hypothesis test, using .30 as the null hypothesis and wanting the power to be (at least) .80 for an assumed true proportion of .75. Would a large or small sample be needed to detect .75 as the true value when the hypothesized value is .30 and a one-sided test is performed? Do the necessary hand computation or use software and determine the necessary sample size, using $\alpha = .05$.

4.3. Explain why it isn't necessary to select a value for $\sigma_{\hat{p}}$ for a one-sample test of a proportion.

4.4. Determine the sample size necessary for detecting a true proportion of .40 in a one-sided test with the null hypothesis being that the proportion is .50. Use $\alpha = .05$.

4.5. Assuming a one-sided test with $\alpha = .05$, determine the sample size necessary to detect an increase in the Poisson parameter from 5 to 6 with a power of .80.

4.6. Show by using the letters for Eq. (4.8) in Eq. (4.5) that the former is equal to the square of the latter.