Suppose that a study is to be carried out in attempt to replicate the results of a previous study in which the odds ratio was found to be ω. If, in the community in which the new study is to be conducted, the rate of occurrence of the event in the first group is P_1, and if the same value ω for the odds ratio is hypothesized to apply in the new community, then the value hypothesized for P_2 is

$$P_2 = \frac{\omega P_1}{\omega P_1 + Q_1}. \tag{4.2}$$

For example, suppose that the value $\omega = 2.5$ had previously been found as the ratio of the odds for depression among female mental hospital patients aged 20–49 to the odds for male mental hospital patients similarly aged. If the same value for the odds ratio is hypothesized to obtain in the mental hospitals of a new community, and if in that community's mental hospitals the rate of depression among male patients aged 20–49 is approximately $P_1 = 0.70$, then the rate among female patients aged 20–49 is hypothesized to be approximately

$$P_2 = \frac{2.5 \times 0.70}{2.5 \times 0.70 + 0.30} = 0.85.$$

An important property of the odds ratio to be demonstrated in Chapters 6 and 7 is that the same value should be obtained whether the study is a prospective or retrospective one. This fact may be taken advantage of if an investigator wishes to replicate a previous study but alters the research design from, say, a retrospective to a prospective study.

Example 4.1.4. Suppose that a case-control (retrospective) study was conducted in a certain school district. School children with emotional disturbances requiring psychological care were compared with presumably normal children on a number of antecedent characteristics. Suppose it was found that one-quarter of the emotionally disturbed children versus one-tenth of the normal controls had lost (by death, divorce, or separation) at least one parent before age 5. The odds ratio is then, from (4.1),

$$\omega = \frac{0.25 \times 0.90}{0.10 \times 0.75} = 3.0.$$

Suppose that a study of this association is to be conducted prospectively in a new community by following through their school year a sample of children who begin school with both parents alive and at home (group 1) and a sample who begin with at least one parent absent from the home (group 2), with the proportions developing emotional problems being compared. From a survey of available school records, the investigator in the new school district is able to estimate that P_1, the proportion of children beginning school with both parents at home who ultimately develop emotional problems, is $P_1 = 0.05$. If the value $\omega = 3.0$ found in the retrospective study is hypothesized to apply in the new school district, the investigator is effectively hypothesizing a value (see equation 4.2)

$$P_2 = \frac{3.0 \times 0.05}{3.0 \times 0.05 + 0.95} = 0.136,$$

or approximately 15%, as the rate of emotional disturbance during school years among children who have lost at least one parent before age 5.

The methods just illustrated may be of use in generating hypotheses for studies to be carried out within a short time, but are likely to prove inadequate for long-term comparative studies. Halperin et al. (1968) give a model and some numerical results when two long-term therapies are to be compared and when few or no dropouts are expected. When dropouts are likely to occur, the model of Schork and Remington (1967) may be useful for generating hypotheses. If the study calls for the comparison of more than two treatments, or if outcome is measured on a scale with more than two categories, the results of Lachin (1977) should be useful.

4.2. THE MATHEMATICS OF SAMPLE SIZE DETERMINATION

We assume in this section and the next that the sample sizes from the two populations being compared, n_1 and n_2, are equal to a common n. We find the value for the common sample size n so that (1) if in fact there is no difference between the two underlying proportions, then the chance is approximately α of falsely declaring the two proportions to differ, and (2) if in fact the proportions are P_1 and $P_2 \neq P_1$, then the chance is approximately $1 - \beta$ of correctly declaring the two proportions to differ. Since this section only derives the mathematical results on which the values in Table A.4 (described in Section 4.3) are based, it is not essential to the sections that follow.

We begin by deriving the sample size, say n', required in both the groups if we ignore the continuity correction. With n' as a first approximation, we then obtain a formula for the desired sample size per group, n, that is appropriate when the test statistic incorporates the continuity correction.

Suppose that the proportions found in the two samples are p_1 and p_2. The statistic used for testing the significance of their difference is, temporarily ignoring the continuity correction,

$$z = \frac{p_2 - p_1}{\sqrt{2\bar{p}\bar{q}/n'}}, \tag{4.3}$$

where

$$\bar{p} = \tfrac{1}{2}(p_1 + p_2)$$

and

$$\bar{q} = 1 - \bar{p}.$$

To assure that the probability of a Type I error is α, the difference between p_1 and p_2 will be declared significant only if

$$|z| > z_{\alpha/2}, \tag{4.4}$$

where $z_{\alpha/2}$ denotes the value cutting off the proportion $\alpha/2$ in the upper tail of the standard normal curve and $|z|$ is the absolute value of z, always a nonnegative quantity. For example, if $\alpha = 0.05$, then $z_{0.05/2} = z_{0.025} = 1.96$, and the difference is declared significant if either $z > 1.96$ or $z < -1.96$.

If the difference between the underlying proportions is actually $P_2 - P_1$, we wish the chances to be $1 - \beta$ of rejecting the hypothesis, that is, of having the outcome represented in (4.4) actually occur. Thus we must find the value of n' such that, when $P_2 - P_1$ is the difference between the proportions,

$$P\left\{\frac{|p_2 - p_1|}{\sqrt{2\bar{p}\bar{q}/n'}} > z_{\alpha/2}\right\} = 1 - \beta. \tag{4.5}$$

The probability in (4.5) is the sum of two probabilities,

$$1 - \beta = P\left\{\frac{p_2 - p_1}{\sqrt{2\bar{p}\bar{q}/n'}} > z_{\alpha/2}\right\} + P\left\{\frac{p_2 - p_1}{\sqrt{2\bar{p}\bar{q}/n'}} < -z_{\alpha/2}\right\}. \tag{4.6}$$

If P_2 is hypothesized to be greater than P_1, then the second probability on the right-hand side of (4.6)—representing the event that p_2 is appreciably less than p_1—is near zero (see Problem 4.1). Thus we need only find the value of n' such that, when $P_2 - P_1$ is the actual difference,

$$1 - \beta = P\left\{\frac{p_2 - p_1}{\sqrt{2\bar{p}\bar{q}/n'}} > z_{\alpha/2}\right\}. \tag{4.7}$$

The probability in (4.7) cannot yet be evaluated, because the mean and the standard error of $p_2 - p_1$ appropriate when $P_2 - P_1$ is the actual difference have not yet been taken into account. The mean of $p_2 - p_1$ is $P_2 - P_1$, and its standard error is

$$\text{se}(p_2 - p_1) = \sqrt{(P_1Q_1 + P_2Q_2)/n'}, \tag{4.8}$$

where $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$.

The following development of (4.7) can be traced using only simple algebra:

$$\begin{aligned} 1 - \beta &= P\left\{(p_2 - p_1) > z_{\alpha/2}\sqrt{2\bar{p}\bar{q}/n'}\right\} \\ &= P\left\{(p_2 - p_1) - (P_2 - P_1) > z_{\alpha/2}\sqrt{2\bar{p}\bar{q}/n'} - (P_2 - P_1)\right\} \\ &= P\left\{\frac{(p_2 - p_1) - (P_2 - P_1)}{\sqrt{(P_1Q_1 + P_2Q_2)/n'}} > \frac{z_{\alpha/2}\sqrt{2\bar{p}\bar{q}/n'} - (P_2 - P_1)}{\sqrt{(P_1Q_1 + P_2Q_2)/n'}}\right\}. \end{aligned} \tag{4.9}$$

The final probability in (4.9) can be evaluated using tables of the normal distribution, because, when the underlying proportions are P_2 and P_1, the quantity

$$Z = \frac{(p_2 - p_1) - (P_2 - P_1)}{\sqrt{(P_1Q_1 + P_2Q_2)/n'}} \tag{4.10}$$

has, to a good approximation if n' is large, the standard normal distribution.

Let z_β denote the value cutting off the proportion β in the upper tail of the standard normal curve. Then, by the symmetry of the normal curve

$$1 - \beta = P(Z > -z_\beta) \tag{4.11}$$

By matching (4.11) with the last probability of (4.9), we find that the value of n' we seek is the one that satisfies

$$\begin{aligned} z_\beta &= \frac{(P_2 - P_1) - z_{\alpha/2}\sqrt{2\bar{p}\bar{q}/n'}}{\sqrt{(P_1Q_1 + P_2Q_2)/n'}} \\ &= \frac{(P_2 - P_1)\sqrt{n'} - z_{\alpha/2}\sqrt{2\bar{p}\bar{q}}}{\sqrt{P_1Q_1 + P_2Q_2}}. \end{aligned} \tag{4.12}$$

Before presenting the final expression for n', we note that (4.12) is a function not only of P_1 and P_2, which may be hypothesized by the investigator, but also of $\bar{p}\bar{q}$, which is observable only after the study is complete. If n' is fairly large, however, $\bar{p}$ will be close to

$$\bar{P} = \frac{P_1 + P_2}{2}, \tag{4.13}$$

and, more importantly, $\bar{p}\bar{q}$ will be close to $\overline{P}\overline{Q}$, where $\overline{Q} = 1 - \overline{P}$. Therefore, replacing $\sqrt{2\bar{p}\bar{q}}$ in (4.12) by $\sqrt{2\bar{P}\bar{Q}}$ and solving for n', we find

$$n' = \frac{\left(z_{\alpha/2}\sqrt{2\bar{P}\bar{Q}} + z_\beta\sqrt{P_1Q_1 + P_2Q_2}\right)^2}{(P_2 - P_1)^2} \tag{4.14}$$

to be the required sample size from *each* of the two populations being compared when the continuity correction is not employed.

Haseman (1978) found that equation (4.14) gives values that are too low, in the sense that the power of the test based on sample size $n_1 = n_2 = n'$ is less than $1 - \beta$ when P_1 and P_2 are the underlying probabilities. Kramer and Greenhouse (1959) proposed an adjustment to (4.14) based on a double use of the continuity correction, once in the statistic (4.3) and again in the statistic (4.10). Their adjustment, which was tabulated in the first edition of this book, was found by Casagrande, Pike, and Smith (1978b) to result in an overcorrection.

By incorporating the continuity correction only in the test statistic (4.3), the latter authors derived

$$n = \frac{n'}{4}\left(1 + \sqrt{1 + \frac{4}{n'|P_2 - P_1|}}\right)^2 \tag{4.15}$$

as the sample size required in each group to provide, to an excellent degree of approximation, the desired significance level and power. The sample sizes tabulated in Table A.4 (which is different from Table A.3 of the first edition) are based on this formula. The values there agree very well with those tabulated by Casagrande, Pike, and Smith (1978a) and Haseman (1978). Ury and Fleiss (1980) present comparisons with some other formulas.

Levin and Chen (1999) show how the continuity correction can be used, in a logically consistent manner, both in the test statistic (4.3) and again in the statistic (4.10) to obtain (4.14) and (4.15). By a careful analysis of the round-off error created in the normal approximation by discreteness in the exact distribution of the test, they demonstrate that while it may appear that (4.14) and (4.15) result from only one use of the $\frac{1}{2}$ continuity correction, two uses are actually required logically. When the second continuity correction is properly employed in the Z statistic (4.10) used for approximating power, it effectively cancels out the round-off error, and leaves only one correction apparent in the formula. Curiously, when Kramer and Greenhouse (1959) considered use of two continuity corrections, they applied the second correction in the wrong direction, which explains the inaccuracy of their sample sizes. Thus Levin and Chen (1999) settle three puzzles from the first two editions of this book: why the sample size table from the first edition had to be replaced (because of an improper use of the second continuity correction); why the Casagrande, Pike, and Smith (1978b) formula is so accurate (because it is a valid approximation to the exact hypergeometric test procedure); and why it is so though it uses only one (logically inconsistent) application of the continuity correction (because it actually uses two logically consistent corrections when properly derived).

To a remarkable degree of accuracy (especially when n' and $|P_2 - P_1|$ are such that $n'|P_2 - P_1| \geq 4$),

$$n = n' + \frac{2}{|P_2 - P_1|}. \tag{4.16}$$

This result, due to Fleiss, Tytun, and Ury (1980), is useful both in arriving quickly at an estimate of required sample sizes and in estimating the power associated with a study involving prespecified sample sizes. Suppose that one can study no more than a total of $2n$ subjects. If the significance level is α and if the two underlying proportions one is seeking to distinguish are P_1 and P_2, one can invert (4.16) and (4.14) to obtain

$$z_\beta = \frac{|P_2 - P_1|\sqrt{n - \frac{2}{|P_2 - P_1|}} - z_{\alpha/2}\sqrt{2\bar{P}\bar{Q}}}{\sqrt{P_1Q_1 + P_2Q_2}} \tag{4.17}$$

as the equation defining the normal curve deviate corresponding to the power $1 - \beta$ associated with the proposed sample sizes. Table A.1 may then be used to find the power itself.

4.3. USING THE SAMPLE SIZE TABLES

Table A.4 gives the equal sample sizes necessary in each of the two groups being compared for varying values of the hypothesized proportions P_1 and P_2, for varying significance levels ($\alpha = 0.01, 0.02, 0.05, 0.10$, and 0.20), and for varying powers [$1 - \beta = 0.50, 0.65(0.05)0.95$, and 0.99]. The value $1 - \beta = 0.50$ is included not so much because an investigator will intentionally embark on a study for which the chances of success are only 50:50, but rather to help provide a baseline for the minimum sample sizes necessary.

The probability of a Type I error, α, is frequently specified first. If, on the basis of declaring the two proportions to differ significantly, the decision is made to conduct further (possibly expensive) research or to replace a standard form of treatment with a new one, the Type I error is serious and α should be kept small (say, 0.01 or 0.02). If the study is aimed only at adding to the body of published knowledge concerning some theory, then the Type I error is less serious, and α may be increased to 0.05 or 0.10 (the more the published evidence points to a difference, the higher may α safely be set).