CHAPTER 5

One- and Two-Sample Inference

5.1 INTRODUCTION

In Chapter 4 we laid the groundwork for statistical inference. The following steps were involved:

1. Define the population of interest.
2. Specify the parameter(s) of interest.
3. Take a random sample from the population.
4. Make statistical inferences about the parameter(s): (a) estimation; and (b) hypothesis testing.

A good deal of "behind-the-scenes" work was necessary, such as specifying what is meant by a *random* sample, but you will recognize that the four steps above summarize the process. In this chapter we (1) formalize the inferential process by defining pivotal quantities and their uses (Section 5.2); (2) consider normal distributions for which *both* the mean and variance are unknown, which will involve the use of the famous Student t-distribution (Sections 5.3 and 5.4); (3) extend the inferential process to a comparison of two normal populations, including comparison of the variances (Sections 5.5 to 5.7); and (4) finally begin to answer the question frequently asked of statisticians: "How many observations should I take?" (Section 5.9).

5.2 PIVOTAL VARIABLES

In Chapter 4, confidence intervals and tests of hypotheses were introduced in a somewhat ad hoc fashion as inference procedures about population parameters. To be able to make inferences, we needed the sampling distributions of the statistics that estimated the parameters. To make inferences about the mean of a normal distribution (with variance known), we needed to know that the sample mean of a random sample was normally distributed; to make inferences about the variance of a normal distribution, we used the chi-square distribution. A pattern also emerged in the development of estimation and hypothesis testing procedures. We discuss next the unifying scheme. This will greatly simplify our understanding of the statistical procedures, so that attention can be focused on the assumptions and appropriateness of such procedures rather than on understanding the mechanics.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

In Chapter 4, we used basically two quantities in making inferences:

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad \chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

What are some of their common features?

1. Each of these expressions involves *at least* a statistic *and* a parameter for the statistic estimated: for example, s^2 and σ^2 in the second formula.
2. The distribution of the quantity was tabulated in a standard normal table or chi-square table.
3. Distribution of the quantity was not dependent on a value of the parameter. Such a distribution is called a *fixed distribution*.
4. Both confidence intervals and tests of hypotheses were derived from a probability inequality involving either Z or χ^2.

Formally, we define:

Definition 5.1. A *pivotal variable* is a function of statistic(s) and parameter(s) having the same fixed distribution (usually tabulated) for all values of the parameter(s).

The quantities Z and χ^2 are pivotal variables. One of the objectives of theoretical statistics is to develop appropriate pivotal variables for experimental situations that cannot be modeled adequately by existing variables.

In Table 5.1 are listed eight pivotal variables and their use in statistical inference. In this chapter we introduce pivotal variables 2, 5, 6, and 8. Pivotal variables 3 and 4 are introduced in Chapter 6. For each variable, the fixed or tabulated distribution is given as well as the formula for a $100(1-\alpha)\%$ confidence interval. The corresponding test of hypothesis is obtained by replacing the statistic(s) by the hypothesized parameter value(s). The table also lists the assumptions underlying the test. Most of the time, the minimal assumption is that of normality of the underlying observations, or appeal is made to the central limit theorem.

Pivotal variables are used primarily in inferences based on the normal distribution. They provide a methodology for estimation and hypothesis testing. The aim of estimation and hypothesis testing is to make probabilistic statements about parameters. For example, confidence intervals and p-values make statements about parameters that have probabilistic aspects. In Chapters 6 to 8 we discuss inferences that do not depend as explicitly on pivotal variables; however, even in these procedures, the methodology associated with pivotal variables is used; see Figure 5.1.

5.3 WORKING WITH PIVOTAL VARIABLES

We have already introduced the manipulation of pivotal variables in Section 4.7. Table 5.1 summarizes the end result of the manipulations. In this section we again outline the process for the case of one sample from a normal population with the variance known. We have a random sample of size n from a normal population with mean μ and variance σ^2 (known). We start with the basic probabilistic inequality

$$P[z_{\alpha/2} \le Z \le z_{1-\alpha/2}] = 1 - \alpha$$

Table 5.1 Pivotal Variables and Their Use in Statistical Inference

	Pivotal Variable	Assumptions: Model	Assumptions: Other[a]	$100(1-\alpha)\%$ Confidence Interval[b]	Inference/Comments
1.	$\dfrac{\overline{Y}-\mu}{\sigma/\sqrt{n}} = Z$	$N(0, 1)$	(i) and (iii); or (ii)	$\overline{Y} \pm \dfrac{z_*\sigma}{\sqrt{n}}$	μ or $\mu = \mu_1 - \mu_2$ based on paired data $z_* = z_{1-\alpha/2}$
2.	$\dfrac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} = Z$	$N(0, 1)$	(i) and (iii); or (ii)	$(\overline{Y}_1-\overline{Y}_2) \pm z_*\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$	$\mu_1 - \mu_2$ based on independent data $z_* = z_{1-\alpha/2}$
3.	$\dfrac{p-\pi}{\sqrt{p(1-p)/n}} = Z$	$N(0, 1)$	(ii)	$p \pm z_*\sqrt{\dfrac{p(1-p)}{n}}$	π $z_* = z_{1-\alpha/2}$
4.	$\dfrac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{p_1q_1/n_1+p_2q_2/n_2}} = Z$	$N(0, 1)$	(ii)	$(p_1-p_2) \pm z_*\sqrt{\dfrac{p_1q_1}{n_1}+\dfrac{p_2q_2}{n_2}}$	$\pi_1 - \pi_2$ based on independent data $z_* = z_{1-\alpha/2}$ $q_1 = 1-p_1$; $q_2 = 1-p_2$
5.	$\dfrac{\overline{Y}-\mu}{s/\sqrt{n}} = t$	t_{n-1}	(i)	$\overline{Y} \pm \dfrac{t_*s}{\sqrt{n}}$	μ or $\mu = \mu_1 - \mu_2$ based on paired data $t_* = t_{n-1,1-\alpha/2}$
6.	$\dfrac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{s_p\sqrt{1/n_1+1/n_2}} = t$	$t_{n_1+n_2-2}$	(i) and (iv)	$(\overline{Y}_1-\overline{Y}_2) \pm t_*s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$	$\mu_1 - \mu_2$ based on independent data $t_* = t_{n_1+n_2-2,1-\alpha/2}$ $s_p^2 = \dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
7.	$\dfrac{(n-1)s^2}{\sigma^2} = \chi^2$	χ^2_{n-1}	(i)	$\dfrac{(n-1)s^2}{\chi_*^2}, \dfrac{(n-1)s^2}{\chi_{**}^2}$	σ^2 $\chi_*^2 = \chi^2_{n-1,1-\alpha/2}$ $\chi_{**}^2 = \chi^2_{n-1,\alpha/2}$
8.	$\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = F$	F_{n_1-1,n_2-1}	(i)	$\dfrac{s_1^2/s_2^2}{F_*}, \dfrac{s_1^2/s_2^2}{F_{**}}$	$\dfrac{\sigma_1^2}{\sigma_2^2}$ $F_* = F_{n_1-1,n_2-1,1-\alpha/2}$ $F_{**} = F_{n_1-1,n_2-1,\alpha/2}$

[a] Assumptions (other): (i) Observations (for paired data, the differences) are independent, normally distributed; (ii) large-sample result; (iii) variance(s) known; (iv) population variances equal.

[b] To determine the appropriate critical region in a test of hypothesis, replace statistic(s) by hypothesized values of parameter(s).

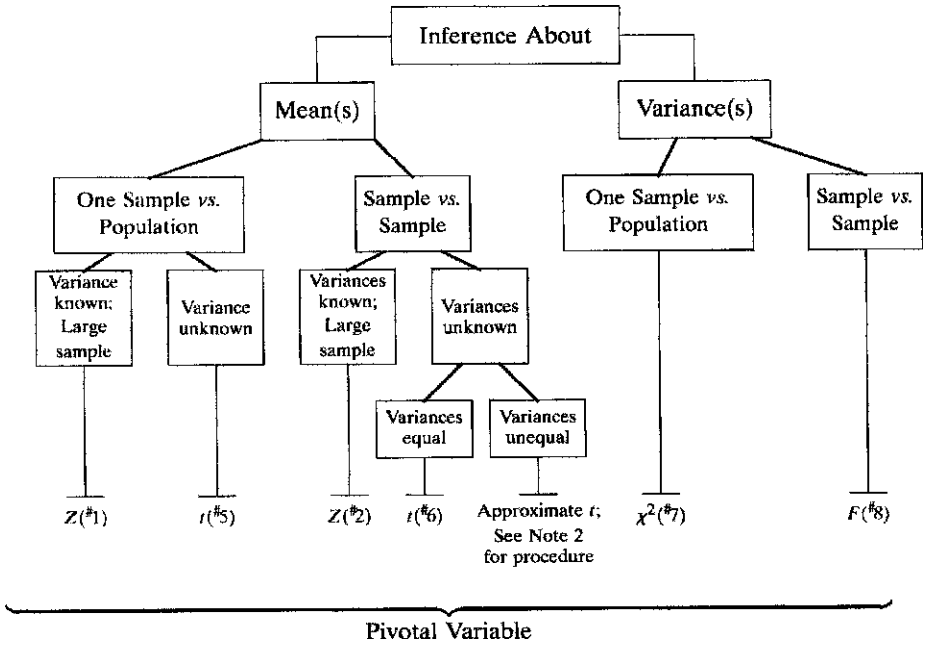


Figure 5.1 Methodology associated with pivotal variables.

We substitute $Z = (\overline{Y} - \mu)/(\sigma_0/\sqrt{n})$, writing σ_0 to indicate that the population variance is assumed to be known:

$$P\left[z_{\alpha/2} \leq \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1 - \alpha$$

Solving for μ produces a $100(1-\alpha)\%$ confidence interval for μ; solving for $\overline{Y}$ and substituting a hypothesized value, μ_0, for μ produces the nonrejection region for a $100(\alpha)\%$ test of the hypothesis:

$100(1 - \alpha)\%$ confidence interval for μ:

$$[\overline{Y} + z_{\alpha/2}\sigma_0/\sqrt{n}, \quad \overline{Y} + z_{1-\alpha/2}\sigma_0/\sqrt{n}]$$

$100(\alpha)\%$ hypothesis test of $\mu = \mu_0$; reject if $\overline{Y}$ is not in

$$[\mu_0 + z_{\alpha/2}\sigma_0/\sqrt{n}, \quad \mu_0 + z_{1-\alpha/2}\sigma_0/\sqrt{n}]$$

Notice again the similarity between the two intervals. These intervals can be written in an abbreviated form using the fact that $z_{\alpha/2} = -z_{1-\alpha/2}$,

$$\overline{Y} \pm \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}} \quad \text{and} \quad \mu_0 \pm \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}$$

for the confidence intervals and tests of hypothesis, respectively.

To calculate the p-value associated with a test statistic, again use is made of the pivotal variable. The null hypothesis value of the parameter is used to calculate the probability of the observed value of the statistic or an observation more extreme. As an illustration, suppose that a population variance is claimed to be $100(\sigma_0^2 = 100)$ vs. a larger value $(\sigma_0^2 > 100)$. From

a random sample of size 11, we are given $s^2 = 220$. What is the p-value for this value (or more extreme)? We use the pivotal quantity $(n-1)s^2/\sigma_0^2$, which under the null hypothesis is chi-square with 10 degrees of freedom.

The one-sided p-value is the probability of a value of $s^2 \geq 220$. Using the pivotal variable, we get

$$P\left[\chi^2 \geq \frac{(11-1)(220)}{100}\right] = P[\chi^2 \geq 22.0]$$

where χ^2 has $11 - 1 = 10$ degrees of freedom, giving a one-sided p-value of 0.0151.

Additional examples in the use of pivotal variables will occur throughout this and later chapters. See Note 5.1 for some additional comments on the pivotal variable approach.

5.4 t-DISTRIBUTION

For a random sample from a normal distribution with mean μ and variance σ^2 (known), the quantity $Z = (\overline{Y} - \mu)/(\sigma/\sqrt{n})$ is a pivotal quantity that has a normal (0,1) distribution. What if the variance is unknown? Suppose that we replace the variance σ^2 by its estimate s^2 and consider the quantity $(\overline{Y} - \mu)/(s/\sqrt{n})$. What is its sampling distribution?

This problem was solved by the statistician W. S. Gossett, in 1908, who published the result under the pseudonym "Student" using the notation

$$t = \frac{\overline{Y} - \mu}{s/\sqrt{n}}$$

The distribution of this variable is now called *Student's t-distribution*. Gossett showed that the distribution of t was similar to that of the normal distribution, but somewhat more "heavy-tailed" (see below), and that for each sample size there is a different distribution. The distributions are indexed by $n - 1$, the degrees of freedom identical to that of the chi-square distribution. The t-distribution is symmetrical, and as the degrees of freedom become infinite, the standard normal distribution is reached.

A picture of the t-distribution for various degrees of freedom, as well as the limiting case of the normal distribution, is given in Figure 5.2. Note that like the standard normal distribution, the t-distribution is bell-shaped and symmetrical about zero. The t-distribution is *heavy-tailed*: The area to the right of a specified positive value is greater than for the normal distribution; in other words, the t-distribution is less "pinched." This is reasonable; unlike a standard normal deviate where only the mean $(\overline{Y})$ can vary (μ and σ are fixed), the t statistic can vary with *both* $\overline{Y}$ and s, so that t will vary even if $\overline{Y}$ is fixed.

Percentiles of the t-distribution are denoted by the symbol $t_{v,\alpha}$, where v indicates the degrees of freedom and α the 100αth percentile. This is indicated in Figure 5.3. In Table 5.1, rather than writing all the subscripts on the t variate, an asterisk is used and explained in the comment part of the table.

Table A.4 lists the percentiles of the t-distribution for each degree of freedom to 30, by fives to 100, and values for 200, 500, and ∞ degrees of freedom. This table lists the t-values such that the percent to the left is as specified by the column heading. For example, for an area of 0.975 (97.5%), the t-value for six degrees of freedom is 2.45. The last row in this column corresponds to a t with an infinite number of degrees of freedom, and the value of 1.96 is identical to the corresponding value of Z; that is, $P[Z \leq 1.96] = 0.975$. You should verify that the last row in this table corresponds precisely to the normal distribution values (i.e., $t_\infty = Z$) and that for practical purposes, t_n and Z are equivalent for $n > 30$. What are the mean and the variance of the t-distribution? The mean will be zero, and the variance is $v/(v-2)$. In the symbols used in Chapter 4, $E(t) = 0$ and $\text{Var}(t) = v/(v-2)$.

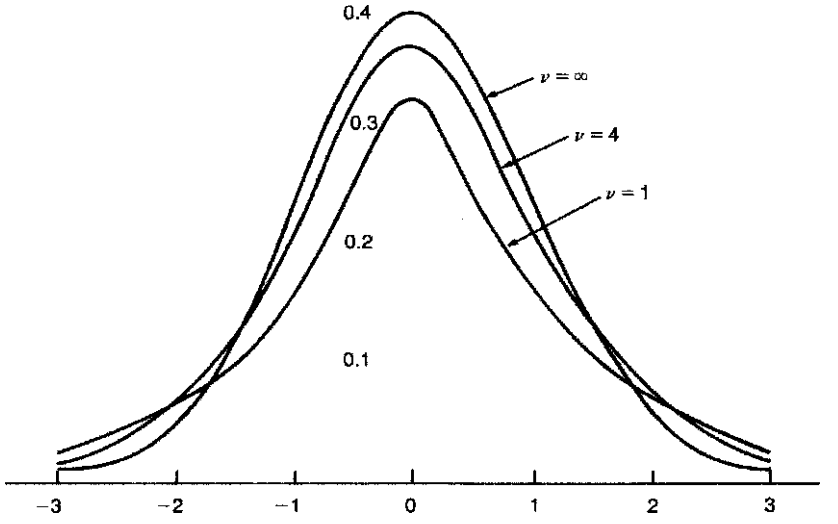


Figure 5.2 Student t-distribution with one, four, and ∞ degrees of freedom.

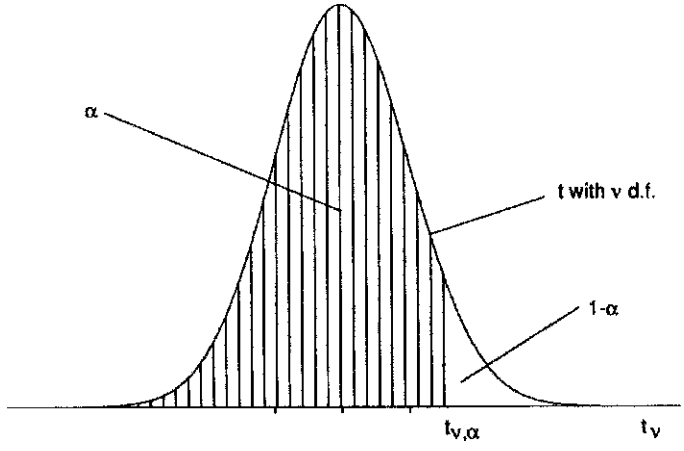


Figure 5.3 Percentiles of the t-distribution.

The converse table of percentiles for a given absolute t-value is given in the Web appendix, and most statistical software will calculate it. We find that the probability of a t-value greater than 1 in absolute value for one degree of freedom is 0.500; the corresponding areas for 7, 30, and ∞ degrees of freedom are 0.351, 0.325, and 0.317, respectively. Thus, at 30 degrees of freedom, the t-distribution is for most practical purposes, indistinguishable from a normal distribution. The term *heavy-tailed* can now be made precise: For a specified value (e.g., with an abscissa value of 1), $P[t_1 \geq 1] > P[t_7 \geq 1] > P[t_{10} \geq 1] > P[Z \geq 1]$.

5.5 ONE-SAMPLE INFERENCE: LOCATION

5.5.1 Estimation and Testing

We begin this section with an example.

Example 5.1. In Example 4.9 we considered the birthweight in grams of the first 11 SIDS cases occurring in King Country in 1969. In this example, we consider the birthweights of the first 15 cases born in 1977. The birthweights for the latter group are

2013	3827	3090	3260	4309	3374	3544	2835
3487	3289	3714	2240	2041	3629	3345	

The mean and standard deviation of this sample are 3199.8 g and 663.00 g, respectively. Without assuming that the population standard deviation is known, can we obtain an interval estimate for the population mean or test the null hypothesis that the population birthweight average of SIDS cases is 3300 g (the same as the general population)?

We can now use the t-distribution. Assuming that birthweights are normally distributed, the quantity

$$\frac{\overline{Y} - \mu}{s/\sqrt{15}}$$

has a t-distribution with $15 - 1 = 14$ degrees of freedom.

Using the estimation procedure, the point estimate of the population mean birthweight of SIDS cases is $3199.8 \doteq 3200$ g. A 95% confidence interval can be constructed on the basis of the t-distribution. For a t-distribution with $15 - 1 = 14$ degrees of freedom, the critical values are ± 2.14, that is, $P[-2.14 \leq t_{14} \leq 2.14] = 0.95$. Using Table 5.1, a 95% confidence interval is constructed using pivotal variable 5:

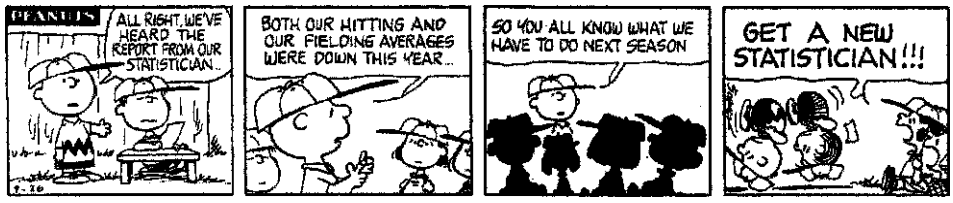
$$3200 \pm \frac{(2.14)(663.0)}{\sqrt{15}} = 3200 \pm 366, \qquad \text{lower limit : 2834 g, upper limit : 3566 g}$$

Several comments are in order:

1. This interval includes 3300 g, the average birthweight in the non-SIDS population. If the analysis had followed a hypothesis-testing procedure, we could not have rejected the null hypothesis on the basis of a two-tailed test.
2. The standard error, $633.0/\sqrt{15}$, is multiplied by 2.14 rather than the critical value 1.96 using a normal distribution. Thus, the confidence interval is wider by approximately 9%. This is the price paid for our ignorance about the value of the population standard deviation. Even in this fairly small sample, the price is modest.

5.5.2 t-Tests for Paired Data

A second example of the one-sample t-test involves its application to paired data. What are *paired data*? Typically, the term refers to repeated or multiple measurements on the same subjects. For example, we may have a measurement of the level of pain before and after administration of an analgesic drug. A somewhat different experiment might consider the level of pain in response to each of *two* drugs. One of these could be a placebo. The first experiment has the weakness that there may be a spontaneous reduction in level of pain (e.g., postoperative pain level), and thus the difference in the responses (after/before) may be made up of two effects: an effect of the drug as well as the spontaneous reduction. Some experimental design considerations are discussed further in Chapter 10. The point we want to make with these two examples is that the basic data consist of pairs, and what we want to look at is the differences within the pairs. If, in the second example, the treatments are to be compared, a common null hypothesis is that the effects are the same and therefore the differences in the treatments should be centered around zero. A natural approach then tests whether the mean of the *sample differences* could have come from a population of differences with mean zero. If we assume that the means of the sample differences are normally distributed, we can apply the t-test (under the null hypothesis),



Cartoon 5.1 PEANUTS. (Reprinted by permission of UFS, Inc.)

Table 5.2 Response of 13 Patients to Aminophylline Treatment at 16 Hours Compared with 24 Hours before Treatment (Apneic Episodes per Hour)

Patient	24 h Before	16 h After	Before–After (Difference)
1	1.71	0.13	1.58
2	1.25	0.88	0.37
3	2.13	1.38	0.75
4	1.29	0.13	1.16
5	1.58	0.25	1.33
6	4.00	2.63	1.37
7	1.42	1.38	0.04
8	1.08	0.50	0.58
9	1.83	1.25	0.58
10	0.67	0.75	−0.08
11	1.13	0.00	1.13
12	2.71	2.38	0.33
13	1.96	1.13	0.83
Total	22.76	12.79	9.97
Mean	1.751	0.984	0.767
Variance	0.7316	0.6941	0.2747
Standard deviation	0.855	0.833	0.524

Source: Data from Bednarek and Roloff [1976].

and estimate the variance of the population of differences σ^2, by the variance of the *sample differences*, s^2.

Example 5.2. The procedure is illustrated with data from Bednarek and Roloff [1976] dealing with the treatment of apnea (a transient cessation of respiration) using a drug, aminophylline, in premature infants. The variable of interest, "average number of apneic episodes per hour," was measured before and after treatment with the drug. An episode was defined as the absence of spontaneous breathing for more than 20 seconds or less if associated with bradycardia or cyanosis.

Patients who had "six or more apneic episodes on each of two consecutive 8 h shifts were admitted to the study." For purposes of the study, consider only the difference between the average number of episodes 24 hours before treatment and 16 hours after. This difference is given in the fourth column of Table 5.2. The average difference for the 13 patients is 0.767 episode per hour. That is, there is a change from 1.751 episodes per hour before treatment to 0.984 episode per hour at 16 hours after treatment.

The standard deviation of the differences is $s = 0.524$. The pivotal quantity to be used is variable 5 from Table 5.1. The argument is as follows: The basic statement about the pivotal variable t with $13 - 1 = 12$ degrees of freedom is $P[-2.18 \leq t_{12} \leq 2.18] = 0.95$ using Table A.4. The form taken for this example is

$$P\left[-2.18 \leq \frac{\overline{Y} - \mu}{0.524/\sqrt{13}} \leq 2.18\right] = 0.95$$

To set up the region to test some hypothesis, we solve for $\overline{Y}$ as before. The region then is

$$P[\mu - 0.317 \leq \overline{Y} \leq \mu + 0.317] = 0.95$$

What is a "reasonable" value to hypothesize for μ? The usual procedure in this type of situation is to assume that the treatment has "no effect." That is, the average difference in the number of apneic episodes from before to after treatment represents random variation. If there is no difference in the population average number of episodes before and after treatment, we can write this as

$$H_0\colon\ \mu = 0$$

We can now set up the hypothesis-testing region as illustrated in Figures 5.4 and 5.5. Figure 5.4 indicates that the sample space can be partitioned without knowing the observed value of $\overline{Y}$. Figure 5.5 indicates the observed value of $\overline{Y} = 0.767$ episode per hour; it clearly falls into the rejection region. Note that the scale has been changed from Figure 5.4 to accommodate the value observed. Hence the null hypothesis is rejected and it is concluded that the average number of apneic episodes observed 16 hours after treatment differs significantly from the average number of apneic episodes observed 24 hours before treatment.

This kind of test is often used when two treatments are applied to the same experimental unit or when the experimental unit is observed over time and a treatment is administered so that it

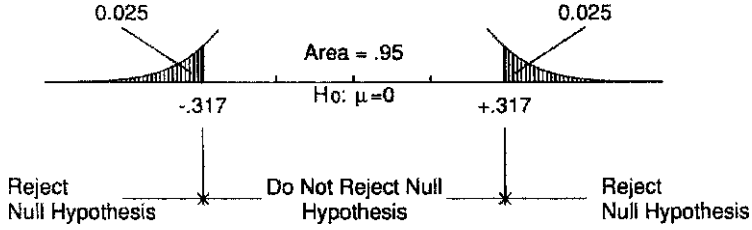


Figure 5.4 Partitioning of sample space of $\overline{Y}$ into two regions: (*a*) region where the null hypothesis is not rejected, and (*b*) region where it is rejected. (Data from Bednarek and Roloff [1976]; see Table 5.2.)

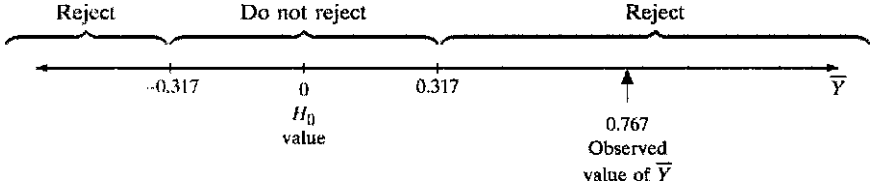


Figure 5.5 Observed value of $\overline{Y}$ and location on the sample space. (Data from Bednarek and Roloff [1976]; see Table 5.2.)

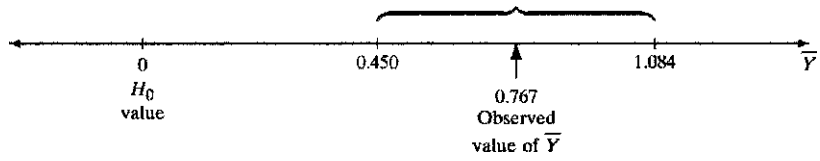


Figure 5.6 A 95% confidence interval for the difference in number of apneic episodes per hour. (Data from Bednarek and Roloff [1976]; see Table 5.2.)

is meaningful to speak of pretreatment and posttreatment situations. As mentioned before, there is the possibility that changes, if observed, are in fact due to changes over time and not related to the treatment.

To construct a confidence interval, we solve the inequality for μ so that we get

$$P[\overline{Y} - 0.317 \leq \mu \leq \overline{Y} + 0.317] = 0.95$$

Again, this interval can be set up to this point without knowing the value of $\overline{Y}$. The value of $\overline{Y}$ is observed to be 0.767 episode per hour, so that the 95% confidence interval becomes

$$[0.767 - 0.317 \leq \mu \leq 0.767 + 0.317] \text{ or } [0.450 \leq \mu \leq 1.084]$$

This interval is displayed in Figure 5.6. Two things should be noted:

1. The width of the confidence interval is the same as the width of the region where the null hypothesis is not rejected (cf. Figure 5.5).
2. The 95% confidence interval does not include zero, the null hypothesis value of μ.

5.6 TWO-SAMPLE STATISTICAL INFERENCE: LOCATION

5.6.1 Independent Random Variables

A great deal of research activity involves the comparison of two or more groups. For example, two cancer therapies may be investigated: one group of patients receives one treatment and a second group the other. The experimental situation can be thought of in two ways: (1) there is one population of subjects, and the treatments induce two subpopulations; or (2) we have two populations that are identical except in their responses to their respective treatments. If the assignment of treatment is random, the two situations are equivalent.

Before exploring this situation, we need to state a definition and a statistical result:

Definition 5.2. Two random variables Y_1 and Y_2 are *statistically independent* if for all fixed values of numbers (say, y_1 and y_2),

$$P[Y_1 \leq y_1, Y_2 \leq y_2] = P[Y_1 \leq y_1]P[Y_2 \leq y_2]$$

The notation $[Y_1 \leq y_1, Y_2 \leq y_2]$ means that Y_1 takes on a value less than or equal to y_1, and Y_2 takes on a value less than or equal to y_2. If we define an event A to have occurred when Y_1 takes on a value less than or equal to y_1, and an event B when Y_2 takes on a value less than or equal to y_2, Definition 5.2 is equivalent to the statistical independence of events $P[AB] = P[A]P[B]$ as defined in Chapter 4. So the difference between statistical independence of random variables and statistical independence of events is that the former in effect describes a relationship between many events (since the definition has to be true for *any* set of values of y_1 and y_2). A basic result can now be stated:

Result 5.1. If Y_1 and Y_2 are statistically independent random variables, then for any two constants a_1 and a_2, the random variable $W = a_1Y_1 + a_2Y_2$ has mean and variance

$$E(W) = a_1E(Y_1) + a_2E(Y_2)$$
$$\text{Var}(W) = a_1^2\text{Var}(Y_1) + a_2^2\text{Var}(Y_2)$$

The only new aspect of this result is that of the variance. In Note 4.10, the expectation of W was already derived. Before giving an example, we also state:

Result 5.2. If Y_1 and Y_2 are statistically independent random variables that are normally distributed, $W = a_1Y_1 + a_2Y_2$ is normally distributed with mean and variance given by Result 5.1.

Example 5.3. Let Y_1 be normally distributed with mean $\mu_1 = 100$ and variance $\sigma_1^2 = 225$; let Y_2 be normally distributed with mean $\mu_2 = 50$ and variance $\sigma_2^2 = 175$. If Y_1 and Y_2 are statistically independent, $W = Y_1 + Y_2$ is normally distributed with mean $100 + 50 = 150$ and variance $225 + 175 = 400$. This and additional examples are given in the following summary:

$$Y_1 \sim N(100, 225),\ Y_2 \sim N(50, 175)$$

W	Mean of W	Variance of W
$Y_1 + Y_2$	150	400
$Y_1 - Y_2$	50	400
$2Y_1 + Y_2$	250	1075
$2Y_1 - 2Y_2$	100	1600

Note that the variance of $Y_1 - Y_2$ is the same as the variance of $Y_1 + Y_2$; this is because the coefficient of Y_1, -1, is squared in the variance formula and $(-1)^2 = (+1)^2 = 1$. In words, the variance of a sum of independent random variables is the same as the variance of a difference of independent random variables.

Example 5.4. Now we look at an example that is more interesting and indicates the usefulness of the two results stated. Heights of females and males are normally distributed with means 162 cm and 178 cm and variances $(6.4\text{ cm})^2$ and $(7.5\text{ cm})^2$, respectively. Let Y_1 = height of female; let Y_2 = height of male. Then we can write

$$Y_1 \sim N(162, (6.4)^2) \quad \text{and} \quad Y_2 \sim N(178, (7.5)^2)$$

Now consider husband–wife pairs. Suppose (probably somewhat contrary to societal *mores*) that husband–wife pairs are formed independent of stature. That is, we interpret this statement to mean that Y_1 and Y_2 are statistically independent. The question is: On the basis of this model, what is the probability that the wife is taller than the husband? We formulate the problem as follows: Construct the new variable $W = Y_1 - Y_2$. From Result 5.2 it follows that

$$W \sim N(-16, (6.4)^2 + (7.5)^2)$$

Now the question can be translated into a question about W; namely, if the wife is taller than the husband, $Y_1 > Y_2$, or $Y_1 - Y_2 > 0$, or $W > 0$. Thus, the question is reformulated as $P[W > 0]$.

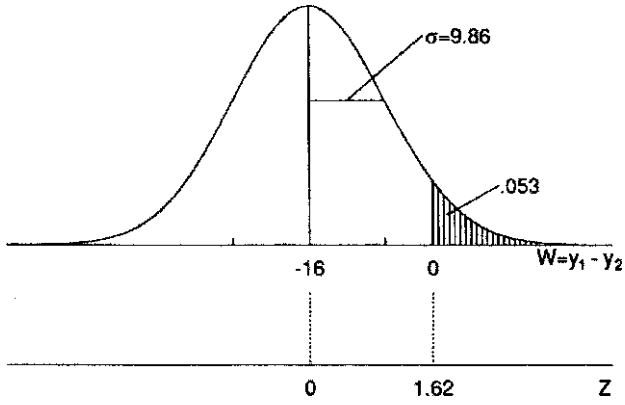


Figure 5.7 Heights of husband–wife pairs.

Hence,

$$\begin{aligned} P[W > 0] &= P\left[Z > \frac{0-(-16)}{\sqrt{(6.4)^2+(7.5)^2}}\right] \\ &\doteq P\left[Z > \frac{16}{9.86}\right] \\ &\doteq P[Z > 1.62] \\ &\doteq 0.053 \end{aligned}$$

so that under the model, in 5.3% of husband–wife pairs the wife will be taller than the husband. Figure 5.7 indicates the area of interest.

5.6.2 Estimation and Testing

The most important application of Result 5.1 involves distribution of the difference of two sample means. If $\overline{Y}_1$ and $\overline{Y}_2$ are the means from two random samples of size n_1 and n_2, respectively, and Y_1 and Y_2 are normally distributed with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, then by Result 5.2,

$$\overline{Y}_1 - \overline{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

so that

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

has a standard normal distribution. This, again, is a pivotal variable, number 2 in Table 5.1. We are now in a position to construct confidence intervals for the quantity $\mu_1 - \mu_2$ or to do hypothesis testing. In many situations, it will be reasonable to assume (null hypothesis) that $\mu_1 = \mu_2$, so that $\mu_1 - \mu_2 = 0$; although the values of the two parameters are unknown, it is reasonable for testing purposes to assume that they are equal, and hence, the difference will be zero. For example, in a study involving two treatments, we could assume that the treatments were equally effective (or ineffective) and that differences between the treatments should be centered at zero.

How do we determine whether or not random variables are statistically independent? The most common way is to note that they are causally independent in the population. That is, the value of Y for one person does not affect the value for another. As long as the observations are sampled independently (e.g., by simple random sampling), they will remain statistically independent. In some situations it is not clear a priori whether variables are independent and there are statistical procedures for testing this assumption. They are discussed in Chapter 9. For the present we will assume that the variables we are dealing with are either statistically independent or if not (as in the case of the paired t-test discussed in Section 5.5.2), use aspects of the data that can be considered statistically independent.

Example 5.5. Zelazo et al. [1972] studied the age at which children walked as related to "walking exercises" given newborn infants. They state that "if a newborn infant is held under his arms and his bare feet are permitted to touch a flat surface, he will perform well-coordinated walking movements similar to those of an adult." This reflex disappears by about eight weeks. They placed 24 white male infants into one of four "treatment" groups. For purposes of this example, we consider only two of the four groups: "active exercise group" and "eight-week control group." The active group received daily stimulation of the walking reflex for eight weeks. The control group was tested at the end of the eight-week treatment period, but there was no intervention. The age at which the child subsequently began to walk was then reported by the mother. The data and basic calculations are shown in Table 5.3.

For purposes of this example, we assume that the sample standard deviations are, in fact, population standard deviations, so that Result 5.2 can be applied. In Example 5.6 we reconsider this example using the two-sample t-test. For this example, we have

$n_1 = 6$	$n_2 = 5$
$\overline{Y}_1 = 10.125$ months	$\overline{Y}_2 = 12.350$ months
$\sigma_1 = 1.4470$ months (assumed)	$\sigma_2 = 0.9618$ month (assumed)

For purposes of this example, the quantity

$$Z = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{(1.4470)^2/6 + (0.9618)^2/5}} = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{0.7307}$$

has a standard normal distribution and is based on pivotal variable 2 of Table 5.1. Let us first set up a 95% confidence interval on the difference $(\mu_1 - \mu_2)$ in the population means. The 95% confidence interval is

$$(\overline{Y}_1 - \overline{Y}_2) \pm 1.96(0.7307)$$

with

$$\text{upper limit} = (10.125 - 12.350) + 1.4322 = -0.79 \text{ month}$$

$$\text{lower limit} = (10.125 - 12.350) - 1.4322 = -3.66 \text{ months}$$

The time line is shown in Figure 5.8.

The 95% confidence interval does not straddle zero, so we would conclude that there is a real difference in age in months when the baby first walked in the exercise group compared to the control group. The best estimate of the difference is $10.125 - 12.350 = -2.22$ months; that is, the age at first walking is about two months earlier than the control group.

Note the flow of the argument: The babies were a homogeneous group before treatment. Allocation to the various groups was on a random basis (assumed but not stated explicitly in the article); the only subsequent differences between the groups were the treatments, so significant differences between the groups must be attributable to the treatments. (Can you think of some reservations that you may want checked before accepting the conclusion?)

Table 5.3 Distribution of Ages (in Months) in Infants for Walking Alone

	Age for Walking Alone	
	Active Exercise Group	Eight-Week Control Group
	9.00	13.25
	9.50	11.50
	9.75	12.00
	10.00	13.50
	13.00	11.50
	9.50	[a]
n	6	5
Mean	10.125	12.350
Standard deviation	1.4470	0.9618

Source: Data from Zelazo et al. [1972].

[a] One observation is missing from the paper.

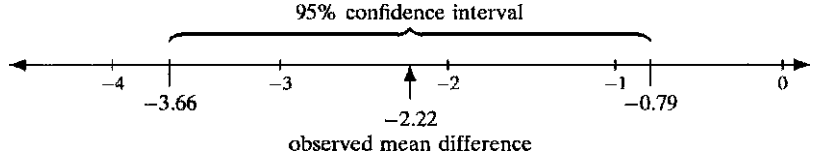


Figure 5.8 Time line for difference in time to infants walking alone.

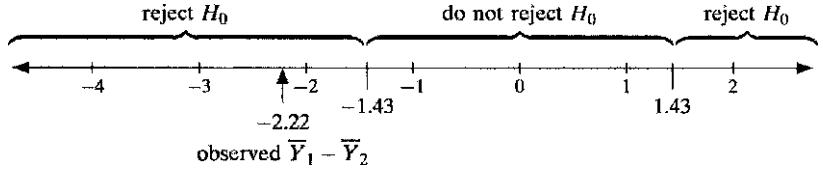


Figure 5.9 Plot showing the nonrejection region.

Formulating the problem as a hypothesis-testing problem is done as follows: A reasonable null hypothesis is that $\mu_1 - \mu_2 = 0$; in this case, the hypothesis of no effect. Comparable to the 95% confidence interval, a test at the 5% level will be carried out. Conveniently, $\mu_1 - \mu_2 = 0$, so that the nonrejection region is simply $0 \pm 1.96(0.7307)$ or 0 ± 1.4322. Plotting this on a line, we get Figure 5.9.

We would reject the null hypothesis, $H_0 : \mu_1 - \mu_2 = 0$, and accept the alternative hypothesis, $H_A : \mu_1 \neq \mu_2$; in fact, on the basis of the data, we conclude that $\mu_1 < \mu_2$.

To calculate the (one-sided) p-value associated with the difference observed, we again use the pivotal variable

$$P\left[[\overline{Y}_1 - \overline{Y}_2] \leq -2.225\right] \doteq P\left[Z \leq \frac{-2.225 - 0}{0.7307}\right]$$

$$\doteq P[Z \leq -3.05]$$

$$\doteq 0.0011$$

The p-value is 0.0011, much less than 0.05, and again, we would reject the null hypothesis. To make the p-value comparable to the two-sided confidence and hypothesis testing procedure, we must multiply it by 2, to give a p-value

$$p\text{-value} = 2(0.0011) = 0.0022$$

We conclude this section by considering the two sample location problem when the population variances are not known. For this we need:

Result 5.3. If $\overline{Y}_1$ and $\overline{Y}_2$ are based on two independent random samples of size n_1 and n_2 from two normal distributions with means μ_1 and μ_2 and the same variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}}$$

has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Here s_p^2 is "the pooled estimate of common variance σ^2," as defined below.

This result is summarized by pivotal variable 6 in Table 5.1. Result 5.3 assumes that the population variances are the same, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. There are then two estimates of σ^2 : s_1^2 from the first sample and s_2^2 from the second sample. How can these estimates be combined to provide the best possible estimate of σ^2? If the sample sizes, n_1 and n_2, differ, the variance based on the larger sample should be given more weight; the pooled estimate of σ^2 provides this. It is defined by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If $n_1 = n_2$, then $s_p^2 = \frac{1}{2}(s_1^2 + s_2^2)$, just the arithmetic average of the variances. For $n_1 \neq n_2$, the variance with the larger sample size receives more weight. See Note 5.2 for a further discussion.

Example 5.5. (*continued*) Consider again the data in Table 5.3 on the age at which children first walk. We will now take the more realistic approach by treating the standard deviations as sample standard deviations, as they should be.

The pooled estimate of the (assumed) common variance is

$$s_p^2 \doteq \frac{(6-1)(1.4470)^2 + (5-1)(0.9618)^2}{6+5-2} \doteq \frac{14.1693}{9} \doteq 1.5744$$

$$s_p \doteq 1.2547 \text{ months}$$

A 95% confidence interval for the difference $\mu_1 - \mu_2$ is constructed first. From Table A.4, the critical t-value for nine degrees of freedom is $t_{9,0.975} = 2.26$. The 95% confidence interval is calculated to be

$$(10.125 - 12.350) \pm (2.26)(1.2547)\sqrt{1/6 + 1/5} \doteq -2.225 \pm 1.717$$

$$\text{lower limit} = -3.94 \text{ months and upper limit} = -0.51 \text{ month}$$

Notice that these limits are wider than the limits $(-3.66, -0.79)$ calculated on the assumption that the variances are known. The wider limits are the price for the additional uncertainty.

The same effect is observed in testing the null hypothesis that $\mu_1 - \mu_2 = 0$. The rejection region (Figure 5.10), using a 5% significance level, is outside

$$0 \pm (2.26)(2.2547)\sqrt{1/6 + 1/5} \doteq 0 \pm 1.72$$

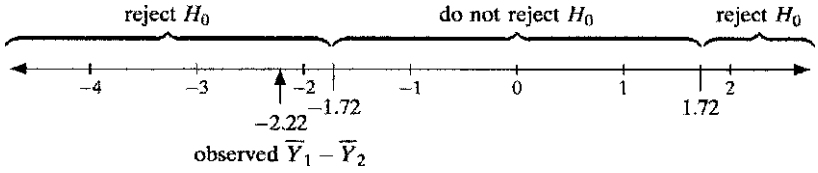


Figure 5.10 Plot showing the rejection region.

The observed value of −2.22 months also falls in the rejection region. Compared to the regions constructed when the variances were assumed known, the region where the null hypothesis is *not* rejected in this case is wider.

5.7 TWO-SAMPLE INFERENCE: SCALE

5.7.1 *F*-Distribution

The final inference procedure to be discussed in this chapter deals with the equality of variances of two normal populations.

Result 5.4. Given two random samples of size n_1 and n_2, with sample variances s_1^2 and s_2^2, from two normal populations with variances σ_1^2 and σ_2^2, the variable

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has an F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

The F-distribution (named in honor of Sir R. A. Fisher) does not have a simple mathematical formula, but most statistical packages can compute tables of the distribution. The F-distribution is indexed by the degrees of freedom associated with s_1^2 (the numerator degrees of freedom) and the degrees of freedom associated with s_2^2 (the denominator degrees of freedom). A picture of the F-distribution is presented in Figure 5.11. The distribution is skewed; the extent of skewness depends on the degrees of freedom. As *both* increase, the distribution becomes more symmetric.

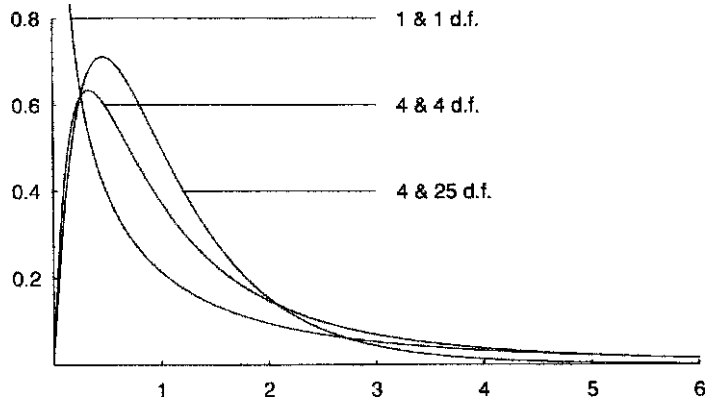


Figure 5.11 F-distribution for three sets of degrees of freedom.

We write $F_{v_1,v_2,\alpha}$ to indicate the 100αth percentile value of an F-statistic with v_1 and v_2 degrees of freedom. The mean of an F-distribution is $v_2/(v_2-2)$, for $v_2 > 2$; the variance is given in Note 5.3. In this note you will also find a brief discussion of the relationship between the four distributions we have now discussed: normal, chi-square, Student t, and F.

It is clear that

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

is a pivotal variable, listed as number 8 in Table 5.1. Inferences can be made on the *ratio* σ_1^2/σ_2^2. [To make inferences about σ_1^2 (or σ_2^2) by itself, we would use the chi-square distribution and the procedure outlined in Chapter 4.] Conveniently, if we want to test whether the variances are equal, that is, $\sigma_1^2 = \sigma_2^2$, the ratio σ_1^2/σ_2^2 is equal to 1 and "drops out" of the pivotal variable, which can then be written

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2}{s_2^2}$$

We would reject the null hypothesis of equality of variances if the observed ratio s_1^2/s_1^2 is "very large" or "very small," how large or small to be determined by the F-distribution.

5.7.2 Testing and Estimation

Continuing Example 5.5, the sample variances in Table 5.3 were $s_1^2 = (1.4470)^2 = 2.0938$ and $s_2^2 = (0.9618)^2 = 0.9251$. Associated with s_1^2 are $6-1=5$ degrees of freedom, and with s_2^2, $5-1=4$ degrees of freedom. Under the null hypothesis of equality of population variances, the ratio s_1^2/s_2^2 has an F-distribution with (5, 4) degrees of freedom. For a two-tailed test at the 10% level, we need $F_{5,4,0.05}$ and $F_{5,4,0.95}$. From Table A.7, the value for $F_{5,4,0.95}$ is 6.26. Using the relationship $F_{v_1,v_2,\alpha} = 1/F_{v_2,v_1,1-\alpha}$, we obtain $F_{5,4,0.05} = 1/F_{4,5,0.95} = 0.19$. The value of F observed is $F_{5,4} = s_1^2/s_2^2 = 2.0938/0.9251 \doteq 2.26$.

From Figure 5.12 it is clear that the null hypothesis of equality of variances is not rejected. Notice that the rejection region is not symmetric about 1, due to the zero bound on the left-hand side. It is instructive to consider F-ratios for which the null hypothesis would have been rejected. On the right-hand side, $F_{5,4,0.95} = 6.26$; this implies that s_1^2 must be 6.26 times as large as s_2^2 before the 10% significance level is reached. On the left-hand side, $F_{5,4,0.05} = 0.19$, so that s_1^2 must be 0.19 times as small as s_2^2 before the 10% significance level is reached. These are reasonably wide limits (even at the 10% level).

At one time statisticians recommended performing an F-test for equality of variances before going on to the t-test. This is no longer thought to be useful. In small samples the F-test cannot reliably detect even quite large differences in variance; in large samples it will reject the hypothesis of equality of variances for differences that are completely unimportant. In addition, the F-test is extremely sensitive to the assumption of normality, even in large samples. The modern solution is to use an approximate version of the t-test that does not assume equal variances (see Note 5.2). This test can be used in all cases or only in cases where the sample variances appear substantially different. In large samples it reduces to the Z-test based on pivotal variable 2 in Table 5.1. The F-test should be restricted to the case where there is a genuine scientific interest in whether two variances are equal.

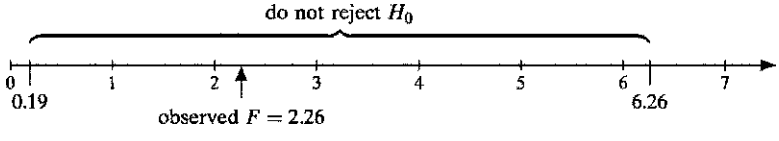


Figure 5.12 Plot showing nonrejection of the null hypothesis of equality of variances.

A few comments about terminology: Sample variances that are (effectively) the same are called *homogeneous*, and those that are not are called *heterogeneous*. A test for equality of population variances, then, is a test for homogeneity or heterogeneity. In the more technical statistical literature, you will find the equivalent terms *homoscedasticity* and *heteroscedasticity tests*.

A confidence interval on the ratios of the population variances σ_1^2/σ_2^2 can be constructed using the pivotal variable approach once more. To set up a $100(1-\alpha)\%$ confidence interval, we need the $100(\alpha/2)$ percentile and $100(1-\alpha/2)$ percentile of the F-distribution.

Continuing with Example 5.5, suppose that we want to construct a 90% confidence interval on σ_1^2/σ_2^2 on the basis of the observed sample. Values for the 5th and 95th percentiles have already been obtained: $F_{5,4,0.05} = 0.19$ and $F_{5,4,0.95} = 6.26$. A 90% confidence interval on σ_1^2/σ_2^2 is then determined by

$$\left(\frac{s_1^2/s_2^2}{F_{5,4,0.95}}, \frac{s_1^2/s_2^2}{F_{5,4,0.05}} \right)$$

For the data observed, this is

$$\left(\frac{2.0938/0.9251}{6.26}, \frac{2.0938/0.9251}{0.19} \right) = (0.36, 11.9)$$

Thus, on the basis of the data observed, we can be 90% confident that the interval (0.36, 11.9) straddles or covers the ratio σ_1^2/σ_2^2 of the population variances. This interval includes 1.0. So, also on the basis of the estimation procedure, we conclude that $\sigma_1^2/\sigma_2^2 = 1$ is not unreasonable.

A 90% confidence interval on the ratio of the standard deviations, σ_1/σ_2, can be obtained by taking square roots of the points (0.36, 11.9), producing (0.60, 3.45) for the interval.

5.8 SAMPLE-SIZE CALCULATIONS

One of the questions most frequently asked of a statistician is: How big must my n be? Stripped of its pseudojargon, a valid question is being asked: How many observations are needed in this study? Unfortunately, the question cannot be answered before additional information is supplied. We first put the requirements in words in the context of a study comparing two treatments; then we introduce the appropriate statistical terminology. To determine sample size, you need to specify or know:

1. How variable the data are
2. The chance that you are willing to tolerate concluding incorrectly that there is an effect when the treatments are equivalent
3. The magnitude of the effect to be detected
4. The certainty with which you wish to detect the effect

Each of these considerations is clearly relevant. The more variation in the data, the more observations are needed to pin down a treatment effect; when there is no difference, there is a chance that a difference will be observed, which due to sampling variability is declared significant. The more certain you want to be of detecting an effect, the more observations you will need, everything else remaining equal. Finally, if the difference in the treatments is very large, a rather economical experiment can be run; conversely, a very small difference in the treatments will require very large sample sizes to detect.

We now phrase the problem in statistical terms: The model we want to consider involves two normal populations with equal variances σ^2, differing at most in their means, μ_1 and μ_2. To determine the sample size, we must specify:

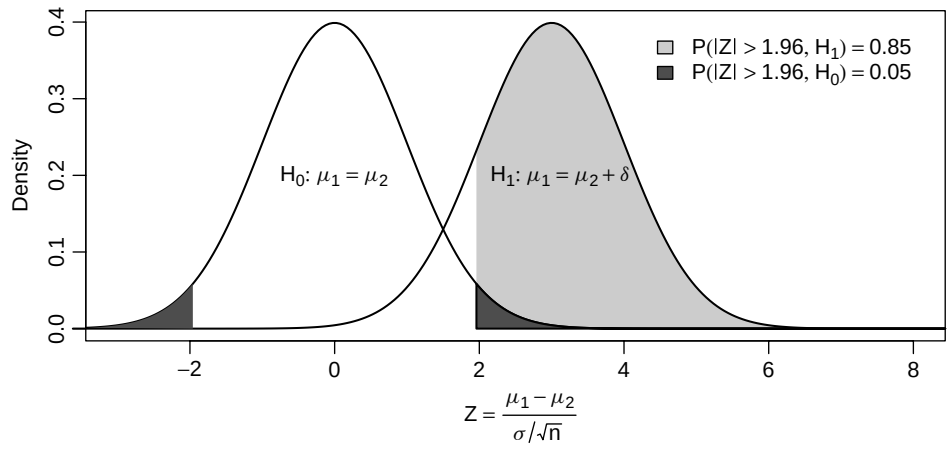


Figure 5.13 Distributions of the Z-statistic under the null and an alternative hypothesis. The probability of $Z < -1.96$ or $Z > 1.96$ under the null hypothesis (the level) is dark gray. The probability under the alternative hypothesis (the power) is light gray.

1. σ^2
2. The probability, α, of a Type I error
3. The magnitude of the difference $\mu_1 - \mu_2$ to be detected
4. The power, $1 - \beta$, or equivalently, the probability of a Type II error, β

Figure 5.13 shows an example of these quantities visually. There are two normal distributions, corresponding to the distribution of the two-sample Z-statistic under the null hypothesis that two means are equal and under the alternative hypothesis that the mean of the first sample is greater than the mean of the second. In the picture, $(\mu_1 - \mu_2)/(\sigma/\sqrt{n}) = 3$, for example, a difference in means of $\mu_1 - \mu_2 = 3$ with a standard deviation $\sigma = 10$ and a sample size of $n = 100$.

The level, or Type I error rate, is the probability of rejecting the null hypothesis if it is true. We are using a 0.05-level two-sided test. The darkly shaded regions are where $Z < -1.96$ or $Z > 1.96$ if $\mu_1 = \mu_2$, adding up to a probability (area under the curve) of 0.05. The power is the probability of rejecting the null hypothesis if it is not true. The lightly shaded region is where $Z > 1.96$ if the alternative hypothesis is true. In theory there is a second lightly shaded region where $Z < -1.96$, but this is invisibly small: There is effectively no chance of rejecting the null hypothesis "in the wrong direction." In this example the lightly shaded region adds up to a probability of 0.85, meaning that we would have 85% power.

Sample sizes are calculated as a function of

$$\Delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

which is defined to be the standardized distance between the two populations. For a two-sided test, the formula for the required sample size *per group* is

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

It is instructive to contemplate this formula. The standardized difference enters as a square. Thus, to detect a treatment different *half* as small as perhaps considered initially will require

four times as many observations per group. Decreasing the probabilities of Type I and Type II errors has the same effect on the sample size; it increases it. However, the increment is not as drastic as it is with Δ. For example, to reduce the probability of a Type I error from 0.05 to 0.025 changes the Z-value from $Z_{0.975} = 1.96$ to $Z_{0.9875} = 2.24$; even though $Z_{1-\alpha/2}$ is squared, the effect will not even be close to doubling the sample size. Finally, the formula assumes that the difference $\mu_1 - \mu_2$ can either be positive or negative. If the direction of the difference can be specified beforehand, a one-tailed value for Z can be used. This will result in a reduction of the sample sizes required for each of the groups, since $z_{1-\alpha}$ would be used.

Example 5.6. At a significance level of $1 - \alpha = 0.95$ (one tail) and power $1 - \beta = 0.80$, a difference $\Delta = 0.3$ is to be detected. The appropriate Z-values are

$$Z_{0.95} = 1.645 \text{ (a more accurate value than given in Table A.2)}$$
$$Z_{0.80} = 0.84$$

The sample size required per group is

$$n = \frac{2(1.645 + 0.84)^2}{(0.3)^2} = 137.2$$

The value is rounded up to 138, so that at least 138 observations *per group* are needed to detect the specified difference at the specified significant level and power.

Suppose that the variance σ^2 is not known, how can we estimate the sample size needed to detect a standardized difference Δ? One possibility is to have an estimate of the variance σ^2 based on a previous study or sample. Unfortunately, no explicit formulas can be given when the variance is estimated; many statistical texts suggest adding between two and four observations per group to get a reasonable approximation to the sample size (see below).

Finally, suppose that *one group*—as in a paired experiment—is to be used to determine whether a populations mean μ differs from a hypothesized mean μ_0. Using the same standardized difference $\Delta = |\mu - \mu_0|/\sigma$, it can be shown that the appropriate number in the group is

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

or one-half the number needed in one group in the two-sample case. This is why tables for sample sizes in the one-sample case tell you, in order to apply the table to the two-sample case, to (1) double the number in the table, and (2) use that number *for each group*.

Example 5.7. Consider data involving PKU children. Assume that IQ in the general population has mean $\mu = 100$ and standard deviation $= 15$. Suppose that a sample of eight PKU children whose diet has been terminated has an average IQ of 94, which is not significantly different from 100. How large would the sample have to be to detect a difference of six IQ points (i.e., the population mean is 94)? The question cannot be answered yet. (Before reading on: What else must be specified?) Additionally, we need to specify the Type I and Type II errors. Suppose that $\alpha = 0.05$ and $\beta = 0.10$. We make the test one-tailed because the alternative hypothesis is that the IQ for PKU children is less than that of children in the general population. A value of $\beta = 0.10$ implies that the power is $1 - \beta = 0.90$. We first calculate the standardized distance

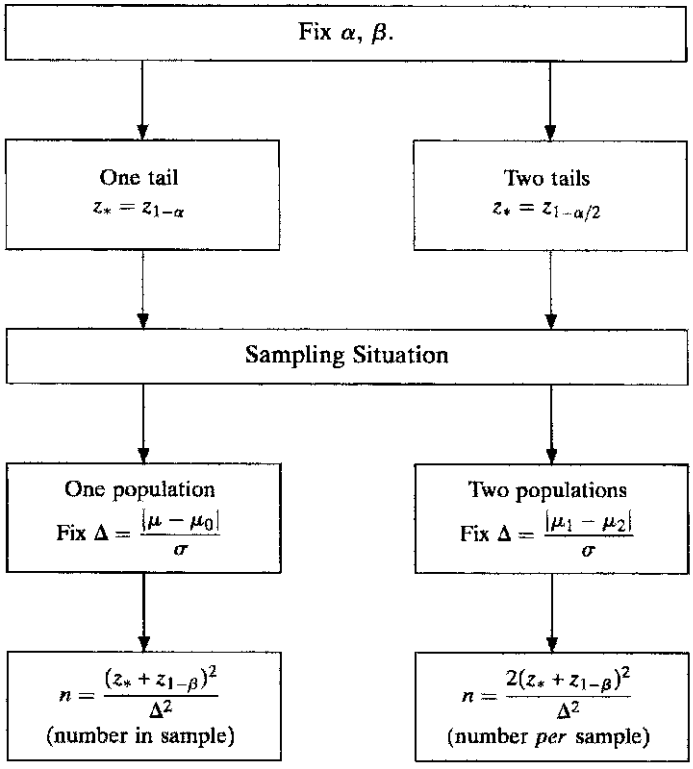
$$\Delta = \frac{|94 - 100|}{15} = \frac{6}{15} = 0.40$$

Then $z_{1-0.05} = z_{0.95} = 1.645$ and $z_{1-0.10} = z_{0.90} = 1.28$. Hence,

$$n = \frac{(1.645 + 1.28)^2}{(0.40)^2} = 53.5$$

Rounding up, we estimate that it will take a sample of 54 observations to detect a difference of $100 - 94 = 6$ IQ points (or greater) with probabilities of Type I and Type II errors as specified.

If the variance is not known, and estimated by s^2, say $s^2 = 15^2$, then statistical tables (not included in this book) indicate that the sample size is 55, not much higher than the 54 we calculated. A summary outline for calculating sample sizes is given in Figure 5.14.



Comments:

1. In the case of two populations, if $\sigma_1^2 \neq \sigma_2^2$, define $\sigma^2 = (\sigma_1^2 + \sigma_2^2)/2$ and proceed as before.
2. If σ is to be estimated from the data, add to the calculated values the following values for an approximate sample size:

		One population	Two populations
One tail	$\alpha = 0.05$	Add 2	Add 1
	$\alpha = 0.01$	Add 4	Add 2
Two tails	$\alpha = 0.05$	Add 2	Add 1
	$\alpha = 0.01$	Add 3	Add 2

Figure 5.14 Sample-size calculations for measurement data.

There is something artificial and circular about all of these calculations. If the difference Δ is known, there is no need to perform an experiment to estimate the difference. Calculations of this type are used primarily to make the researcher aware of the kinds of differences that can be detected. Often, a calculation of this type will convince a researcher *not* to carry out a piece of research, or at least to think very carefully about the possible ways of increasing precision, perhaps even contemplating a radically different attack on the problem. In addition, the size of a sample may be limited by considerations such as cost, recruitment rate, or time constraints beyond control of the investigations. In Chapter 6 we consider questions of sample size for discrete variables.

NOTES

5.1 Inference by Means of Pivotal Variables: Some Comments

1. The problem of finding pivotal variables is a task for statisticians. Part of the problem is that such variables are not always unique. For example, when working with the normal distribution, why not use the sample median rather than the sample mean? After all, the median is admittedly more robust. However, the variance of the sample median is larger than that of the sample mean, so that a less precise probabilistic statement would result.

2. In many situations there is no exactly pivotal variable available in small samples, although a pivotal variable can typically be found in large samples.

3. The principal advantage of using the pivotal variable approach is that it gives you a unified picture of a great number of procedures that you will need.

4. There is a philosophical problem about the interpretation of a confidence interval. For example, consider the probability inequality

$$P[-1.96 \leq Z \leq 1.96] = 0.95$$

which leads to a 95% confidence interval for the mean of a normal population on the basis of a random sample of size n:

$$P\left[\overline{Y} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + \frac{1.96\sigma}{\sqrt{n}}\right] = 0.95$$

It is argued that once $\overline{Y}$ is observed, *this* interval either covers the mean or not; that is, P is either 0 or 1. One answer is that probabilities are not associated with a particular event—whether they have occurred or may occur at some future time—but with a population of events. For this reason we say *after the fact* that we are 95% confident that the mean is in the interval, *not* that the probability is 0.95 that the mean is in the interval.

5. Given two possible values for a parameter, which one will be designated as the null hypothesis value and which one as the alternative hypothesis value in a hypothesis testing situation? If nothing else is given, the designation will be arbitrary. Usually, there are at least four considerations in designating the *null value* of a parameter:

- **a.** Often, the null value of the parameter permits calculation of a p-value. For example, if there are two hypotheses, $\mu = \mu_0$ and $\mu \neq \mu_0$, only under $\mu = \mu_0$ can we calculate the probability of an occurrence of the observed value or a more extreme value.
- **b.** Past experience or previous work may suggest a specified value. The new experimentation or treatment then has a purpose: rejection of the value established previously, or assessment of the magnitude of the change.

c. Occam's razor can be appealed to. It states: "Do not multiply hypotheses beyond necessity," meaning in this context that we usually start from the value of a parameter that we would assume if no new data were available or to be produced.

d. Often, the null hypothesis is a "straw man" we hope to reject, for example, that a new drug has the same effect as a placebo.

6. Sometimes it is argued that the smaller the p-value, the stronger the treatment effect. You now will recognize that this cannot be asserted baldly. Consider the two-sample t-test. A p-value associated with this test will depend on the quantities $\mu_1 - \mu_2, s_p, n_1$, and n_2. Thus, differences in p-values between two experiments may simply reflect differences in sample size or differences in background variability (as measured by s_p).

5.2 Additional Notes on the t-Test

1. *Heterogeneous variances in the two-sample*. t-*test*. Suppose that the assumption of homogeneity of variances in the two-sample t-test is not tenable. What can be done? At least three avenues are open:

a. Use an approximation to the t procedure.

b. Transform the data.

c. Use another test, such as a nonparametric test.

With respect to the last point, alternative approaches are discussed in Chapter 8. With respect to the first point, one possibility is to rank the observations from smallest to largest (disregarding group membership) and then carry out the t-test on the ranks. This is a surprisingly good *test* but does not allow us to estimate the magnitude of the difference between the two groups. See Conover and Iman [1981] for an interesting discussion and Thompson [1991] for some precautions. Another approach adjusts the degrees of freedom of the two-sample t-test. The procedure is as follows: Let $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$, and samples of size n_1 and n_2 are taken, respectively. The variable

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

has a standard normal distribution. However, the analogous quantity with the population variances σ_1^2 and σ_2^2 replaced by the sample variances s_1^2 and s_2^2 does not have a t-distribution. The problem of finding the distribution of this quantity is known as the *Behrens–Fisher problem*. It is of theoretical interest in statistics because there is no exact solution to such an apparently simple problem. There are, however, perfectly satisfactory practical solutions. One approach adjusts the degrees of freedom of this statistic in relation to the extent of dissimilarity of the two sample variances. The t-table is entered not with $n_1 + n_2 - 2$ degrees of freedom, but with

$$\text{degrees of freedom} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 + 1) + (s_2^2/n_2)^2/(n_2 + 1)} - 2$$

This value need not be an integer; if you are working from tables of the t-distribution rather than software, it may be necessary to round down this number. The error in this approximation is very small and is likely to be negligible compared to the errors caused by nonnormality. For

large samples (e.g., $n_1, n_2 > 30$), the statistic

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

can be treated as a standard normal deviate even if the distribution of Y is not a normal distribution.

2. *The two-sample* t*-test and design of experiments*. Given that a group has to be divided into two subgroups, the arrangement that minimizes the standard error of the difference is that of equal sample sizes in each group when there is a common σ^2. To illustrate, suppose that 10 objects are to be partitioned into two groups; consider the multiplier $\sqrt{1/n_1 + 1/n_2}$, which determines the relative size of the standard errors.

n_1	n_2	$\sqrt{1/n_1 + 1/n_2}$
5	5	0.63
6	4	0.65
7	3	0.69
8	2	0.79

This list indicates that small deviations from a 5 : 5 ratio do not affect the multiplier very much. It is sometimes claimed that sample sizes must be equal for a valid t-test: Except for giving the smallest standard error of the difference, there is no such constraint.

3. *The "wrong"* t*-test*. What is the effect of carrying out a two-sample t-test on paired data, that is, data that should have been analyzed by a paired t-test? Usually, the level of significance is reduced. On the one hand, the degrees of freedom are *increased* from $(n - 1)$, assuming n pairs of observations, to $2(n - 1)$, but at the same time, additional variation is introduced, so that the standard error of the difference is now larger. In any event the assumption of statistical independence between "groups" is usually inappropriate.

4. *Robustness of the* t*-test*. The t-test tends to be sensitive to *outliers*, unusually small or large values of a variable. We discuss other methods of analysis in Chapter 8. As a matter of routine, you should always graph the data in some way. A simple box plot or histogram will reveal much about the structure of the data. An outlier may be a perfectly legitimate value and its influence on the t-test entirely appropriate, but it is still useful to know that this influence is present.

5.3 Relationships and Characteristics of the Fixed Distributions in This Chapter

We have already suggested some relationships between the fixed distributions. The connection is more remarkable yet and illustrates the fundamental role of the normal distribution. The basic connection is between the standard normal and the chi-square distribution. Suppose that we draw randomly 10 independent values from a standard normal distribution, square each value, and sum them. This sum is a random variable. What is its sampling distribution? It turns out to be chi-square with 10 degrees of freedom. Using notation, let $Z_1, Z_2, \ldots, Z_{10}$ be the values of Z obtained in drawings 1 to 10. Then, $Z_1^2 + \cdots + Z_{10}^2$ has a chi-square distribution with 10 degrees of freedom: $\chi_{10}^2 = Z_1^2 + \cdots + Z_{10}^2$. This generalizes the special case $\chi_1^2 = Z^2$.

The second connection is between the F-distribution and the chi-square distribution. Suppose that we have two independent chi-square random variables with v_1 and v_2 degrees of freedom. The ratio

$$\frac{\chi_{v_1}^2/v_1}{\chi_{v_2}^2/v_2} = F_{v_1, v_2}$$

has an F-distribution with v_1 and v_2 degrees of freedom. Finally, the square of a t-variable with v degrees of freedom is $F_{1,v}$. Summarizing yields

$$\chi_v^2 = \sum_{i=1}^{v} Z_i^2, \qquad t_v^2 = F_{1,v} = \frac{\chi_1^2/1}{\chi_v^2/v}$$

A special case connects all four pivotal variables:

$$Z^2 = t_\infty^2 = \chi_1^2 = F_{1,\infty}$$

Thus, given the F-table, all the other tables can be generated from it. For completeness, we summarize the mean and variance of the four fixed distributions:

Distribution	Symbol	Mean	Variance	
Normal	Z	0	1	
Student t	t_v	0	$\frac{v}{v-2}$	$(v > 2)$
Chi-square	χ_v^2	v	$2v$	
Fisher's F	F_{v_1,v_2}	$\dfrac{v_2}{v_2-2}$	$\dfrac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$	$(v_2 > 4)$

5.4 One-Sided Tests and One-Sided Confidence Intervals

Corresponding to one-sided (one-tailed) tests are one-sided confidence intervals. A one-sided confidence interval is derived from a pivotal quantity in the same was as a two-sided confidence interval. For example, in the case of a one-sample t-test, a pivotal equation is

$$P\left[-\infty \le \frac{\overline{x}-\mu}{s/\sqrt{n}} \le t_{n-1,1-\alpha}\right] = 1-\alpha$$

Solving for μ produces a $100(1-\alpha)\%$ *upper* one-sided confidence interval for μ : $(\overline{x} - t_{n-1,1-\alpha}s/\sqrt{n}, \infty)$. Similar intervals can be constructed for all the pivotal variables.

PROBLEMS

5.1 Rickman et al. [1974] made a study of changes in serum cholesterol and triglyceride levels of subjects following the Stillman diet. The diet consists primarily of protein and animal fats, restricting carbohydrate intake. The subjects followed the diet with length of time varying from 3 to 17 days. (Table 5.4). The mean cholesterol level increased significantly from 215 mg per/100 mL at baseline to 248 mg per/100 mL at the end of the diet. In this problem, we deal with the triglyceride level.

(a) Make a histogram or stem-and-leaf diagram of the *changes* in triglyceride levels.

(b) Calculate the average change in triglyceride level. Calculate the standard error of the difference.

(c) Test the significance of the average change.

(d) Construct a 90% confidence interval on the difference.

(e) The authors indicate that subjects (5,6), (7,8), (9,10), and (15,16) were "repeaters," that is, subjects who followed the diet for two sequences. Do you think it is

Table 5.4 Diet Data for Problem 5.1

Subject	Days on Diet	Weight (kg)		Triglyceride (mg/100 ml)	
		Initial	Final	Baseline	Final
1	10	54.6	49.6	159	194
2	11	56.4	52.8	93	122
3	17	58.6	55.9	130	158
4	4	55.9	54.6	174	154
5	9	60.0	56.7	148	93
6	6	57.3	55.5	148	90
7	3	62.7	59.6	85	101
8	6	63.6	59.6	180	99
9	4	71.4	69.1	92	183
10	4	72.7	70.5	89	82
11	4	49.6	47.1	204	100
12	7	78.2	75.0	182	104
13	8	55.9	53.2	110	72
14	7	71.8	68.6	88	108
15	7	71.8	66.8	134	110
16	14	70.5	66.8	84	81

reasonable to include their data the "second time around" with that of the other subjects? Supposing not, how would you now analyze the data? Carry out the analysis. Does it change your conclusions?

5.2 In data of Dobson et al. [1976], 36 patients with a confirmed diagnosis of phenylketonuria (PKU) were identified and placed on dietary therapy before reaching 121 days of age. The children were tested for IQ (Stanford–Binet test) between the ages of 4 and 6; subsequently, their normal siblings of closest age were also tested with the Stanford–Binet. The following are the first 15 pairs listed in the paper:

Pair	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
IQ of PKU case	89	98	116	67	128	81	96	116	110	90	76	71	100	108	74
IQ of sibling	77	110	94	91	122	94	121	114	88	91	99	93	104	102	82

(a) State a suitable null and an alternative hypotheses with regard to these data.

(b) Test the null hypothesis.

(c) State your conclusions.

(d) What are your assumptions?

(e) Discuss the concept of power with respect to this set of data using the fact that PKU invariably led to mental retardation until the cause was found and treatment comprising a restricted diet was instituted.

(f) The mean difference (PKU case − sibling) in IQ for the full 36 pairs was −5.25; the standard deviation of the difference was 13.18. Test the hypothesis of no difference in IQ for this entire set of data.

5.3 Data by Mazze et al. [1971] deal with the preoperative and postoperative creatinine clearance (ml/min) of six patients anesthetized by halothane:

	Patient					
	1	2	3	4	5	6
Preoperative	110	101	61	73	143	118
Postoperative	149	105	162	93	143	100

(a) Why is the paired t-test preferable to the two-sample t-test in this case?

(b) Carry out the paired t-test and test the significance of the difference.

(c) What is the model for your analysis?

(d) Set up a 99% confidence interval on the difference.

(e) Graph the data by plotting the pairs of values for each patient.

5.4 Some of the physiological effects of alcohol are well known. A paper by Squires et al. [1978] assessed the acute effects of alcohol on auditory brainstem potentials in humans. Six volunteers (including the three authors) participated in the study. The latency (delay) in response to an auditory stimulus was measured before and after an intoxicating dose of alcohol. Seven different peak responses were identified. In this exercise, we discuss only latency peak 3. Measurements of the latency of peak (in milliseconds after the stimulus onset) in the six subjects were as follows:

	Latency of Peak					
	1	2	3	4	5	6
Before alcohol	3.85	3.81	3.60	3.68	3.78	3.83
After alcohol	3.82	3.95	3.80	3.87	3.88	3.94

(a) Test the significance of the difference at the 0.05 level.

(b) Calculate the p-value associated with the result observed.

(c) Is your p-value based on a one- or two-tailed test? Why?

(d) As in Problem 5.3, graph these data and state your conclusion.

(e) Carry out an (incorrect) two-sample test and state your conclusions.

(f) Using the sample variances s_1^2 and s_2^2 associated with the set of readings observed before and after, calculate the variance of the difference, *assuming* independence (call this variance 1). How does this value compare with the variance of the difference calculated in part (a)? (Call this variance 2.) Why do you suppose variance 1 is so much bigger than variance 2? The *average* of the differences is the same as the difference in the averages. Show this. Hence, the two-sample t-test differed from the paired t-test only in the divisor. Which of the two tests in more powerful in this case, that is, declares a difference significant when in fact there is one?

5.5 The following data from Schechter et al. [1973] deal with sodium chloride preference as related to hypertension. Two groups, 12 normal and 10 hypertensive subjects, were isolated for a week and compared with respect to Na^+ intake. The following are the average daily Na^+ intakes (in milligrams):

Normal	10.2	2.2	0.0	2.6	0.0	43.1	45.8	63.6	1.8	0.0	3.7	0.0
Hypertensive	92.8	54.8	51.6	61.7	250.8	84.5	34.7	62.2	11.0	39.1		

(a) Compare the average daily Na^+ intake of the hypertensive subjects with that of the normal volunteers by means of an appropriate t-test.

(b) State your assumptions.

(c) Assuming that the population variances are not homogeneous, carry out an appropriate t-test (see Note 5.2).

5.6 Kapitulnik et al. [1976] compared the metabolism of a drug, zoxazolamine, in placentas from 13 women who smoked during pregnancy and 11 who did not. The purpose of the study was to investigate the presence of the drug as a possible proxy for the rate at which benzo[a]pyrene (a by-product of cigarette smoke) is metabolized. The following data were obtained in the measurement of zoxazolamine hydroxylase production (nmol $3H_2O$ formed/g per hour):

Nonsmoker	0.18	0.36	0.24	0.50	0.42	0.36	0.50	0.60	0.56	0.36	0.68		
Smoker	0.66	0.60	0.96	1.37	1.51	3.56	3.36	4.86	7.50	9.00	10.08	14.76	16.50

(a) Calculate the sample mean and standard deviation for each group.

(b) Test the assumption that the two sample variances came from a population with the same variance.

(c) Carry out the t-test using the approximation to the t-procedure discussed in Note 5.2. What are your conclusions?

(d) Suppose we agree that the variability (as measured by the standard deviations) is proportional to the level of the response. Statistical theory then suggests that the logarithms of the responses should have roughly the same variability. Take logarithms of the data and test, once more, the homogeneity of the variances.

5.7 Sometime you may be asked to do a two-sample t-test knowing only the mean, standard deviation, and sample sizes. A paper by Holtzman et al. [1975] dealing with terminating a phenylalanine-restricted diet in 4-year-old children with phenylketonuria (PKU) illustrates the problem. The purpose of the diet is to reduce the phenylalanine level. A high level is associated with mental retardation. After obtaining informed consent, eligible children of 4 years of age were randomly divided into two groups. Children in one group had their restricted diet terminated while children in the other group were continued on the restricted diet. At 6 years of age, the phenylalanine levels were tested in all children and the following data reported:

	Diet Terminated	**Diet Continued**
Number of children	5	4
Mean phenylalanine level (mg/dl)	26.9	16.7
Standard deviation	4.1	7.3

(a) State a reasonable null hypothesis and alternative hypothesis.

(b) Calculate the pooled estimate of the variance s_p^2.

(c) Test the null hypothesis of part (a). Is your test one-tailed, or two? Why?

(d) Test the hypothesis that the sample variances came from two populations with the same variance.

(e) Construct a 95% confidence interval on the difference in the population phenylalanine levels.

(f) Interpret the interval constructed in part (e).

(g) "This set of data has little power," someone says. What does this statement mean? Interpret the implications of a Type II error in this example.

(h) What is the interpretation of a Type I error in this example? Which, in your opinion, is more serious in this example: a Type I error or a Type II error?

(i) On the basis of these data, what would you recommend to a parent with a 4-year-old PKU child?

(j) Can you think of some additional information that would make the analysis more precise?

5.8 Several population studies have demonstrated an inverse correlation of sudden infant death syndrome (SIDS) rate with birthweight. The occurrence of SIDS in one of a pair of twins provides an opportunity to test the hypothesis that birthweight is a major determinant of SIDS. The data shown in Table 5.5 consist of the birthweights (in grams) of each of 22 dizygous twins and each of 19 monozygous twins.

(a) With respect to the dizygous twins, test the hypothesis given above. State the null hypothesis.

(b) Make a similar test on the monozygous twins.

(c) Discuss your conclusions.

Table 5.5 Birthweight Data for Problem 5.8

Dizygous Twins		Monozygous Twins	
SID	Non-SID	SID	Non-SID
1474	2098	1701	1956
3657	3119	2580	2438
3005	3515	2750	2807
2041	2126	1956	1843
2325	2211	1871	2041
2296	2750	2296	2183
3430	3402	2268	2495
3515	3232	2070	1673
1956	1701	1786	1843
2098	2410	3175	3572
3204	2892	2495	2778
2381	2608	1956	1588
2892	2693	2296	2183
2920	3232	3232	2778
3005	3005	1446	2268
2268	2325	1559	1304
3260	3686	2835	2892
3260	2778	2495	2353
2155	2552	1559	2466
2835	2693		
2466	1899		
3232	3714		

Source: D. R. Peterson, Department of Epidemiology, University of Washington.

5.9 A pharmaceutical firm claims that a new analgesic drug relieves mild pain under standard conditions for 3 hours with a standard deviation of 1 hour. Sixteen patients are tested under the same conditions and have an average pain relief of 2.5 hours. The hypothesis that the population mean of this sample is also 3 hours is to be tested against the hypothesis that the population mean is in fact less than 3 hours; $\alpha = 0.5$.

(a) What is an appropriate test?

(b) Set up the appropriate critical region.

(c) State your conclusion.

(d) Suppose that the sample size is doubled. State precisely how the nonrejection region for the null hypothesis is changed.

5.10 Consider Problem 3.9, dealing with the treatment of essential hypertension. Compare treatments A and B by means of an appropriate t-test. Set up a 99% confidence interval on the reduction of blood pressure under treatment B as compared to treatment A.

5.11 During July and August 1976, a large number of Legionnaires attending a convention died of mysterious and unknown cause. Epidemiologists talked of "an outbreak of Legionnaires' disease." One possible cause was thought to be toxins: nickel, in particular. Chen et al. [1977] examined the nickel levels in the lungs of nine of the cases, and selected nine controls. All specimens were coded by the Centers for Disease Control in Atlanta before being examined by the investigators. The data are as follows (μg per 100 g dry weight):

Legionnaire cases	65	24	52	86	120	82	399	87	139
Control cases	12	10	31	6	5	5	29	9	12

Note that there was no attempt to match cases and controls.

(a) State a suitable null hypothesis and test it.

(b) We now know that Legionnaires' disease is caused by a bacterium, genus *Legionella*, of which there are several species. How would you explain the "significant" results obtained in part (a)? (Chen et al. [1977] consider various explanations also.)

5.12 Review Note 5.3. Generate a few values for the normal, t, and chi-square tables from the F-table.

5.13 It is claimed that a new drug treatment can substantially reduce blood pressure. For purposes of this exercise, assume that only diastolic blood pressure is considered. A certain population of hypertensive patients has a mean blood pressure of 96 mmHg. The standard deviation of diastolic blood pressure (variability from subject to subject) is 12 mmHg. To be biologically meaningful, the new drug treatment should lower the blood pressure to at least 90 mmHg. A random sample of patients from the hypertensive population will be treated with the new drug.

(a) Assuming that $\alpha = 0.05$ and $\beta = 0.05$, calculate the sample size required to demonstrate the effect specified.

(b) Considering the labile nature of blood pressure, it might be argued that any "treatment effect" will merely be a "put-on-study effect." So the experiment is redesigned to consider two random samples from the hypertensive population, one of which will receive the new treatment, and the other, a placebo. Assuming the same specifications as above, what is the required sample size per group?

(c) Blood pressure readings are notoriously variable. Suppose that a subject's diastolic blood pressure varies randomly from measurement period to measurement period with a standard deviation of 4 mmHg. Assuming that measurement variability is independent of subject-to-subject variability, what is the overall variance or the total variability in the population? Recalculate the sample sizes for the situation described in parts (a) and (b).

(d) Suppose that the *change* in blood pressure from baseline is used. Suppose that the standard deviation of the change is 6 mmHg. How will this change the sample sizes of parts (a) and (b)?

5.14 In a paper in the *New England Journal of Medicine*, Rodeheffer et al. [1983] assessed the effect of a medication, nifedipine, on the number of painful attacks in patients with Raynaud's phenomenon. This phenomenon causes severe digital pain and functional disability, particularly in patients with underlying connective tissue disease. The drug causes "vascular smooth-muscle relaxation and relief of arterial vasospasm." In this study, 15 patients were selected and randomly assigned to one of two treatment sequences: placebo–nifedipine, or nifedipine–placebo. The data in Table 5.6 were obtained.

(a) Why were patients *randomly* assigned to one of the two sequences? What are the advantages?

(b) The data of interest are in the columns marked "placebo" and "nifedipine." State a suitable null hypothesis and alternative hypothesis for these data. Justify your choices. Test the significance of the difference in total number of attacks in two weeks on placebo with that of treatment. Use a t-test on the differences in the response. Calculate the p-value.

(c) Construct a 95% confidence interval for the difference. State your conclusions.

(d) Make a scatter plot of the placebo response (x-axis) vs. the nifedipine response (y-axis). If there was no significant difference between the treatments, about what line should the observations be scattered?

(e) Suppose that a statistician considers only the placebo readings and calculates a 95% confidence interval on the population mean. Similarly, the statistician calculates a 95% confidence interval on the nifedipine mean. A graph is made to see if the intervals overlap. Do this for these data. Compare your results with that of part (c). Is there a contradiction? Explain.

(f) One way to get rid of outliers is to carry out the following procedure: Take the differences of the data in columns 7 (placebo) and 9 (nifedipine), and rank them disregarding the signs of the differences. Put the sign of the difference on the rank. Now, carry out a paired t-test on the signed ranks. What would be an appropriate null hypothesis? What would be an appropriate alternative hypothesis? Name one advantage and one disadvantage of this procedure. (It is one form of a nonparametric test discussed in detail in Chapter 8.)

5.15 Rush et al. [1973] reported the design of a randomized controlled trial of nutritional supplementation in pregnancy. The trial was to be conducted in a poor American black population. The variable of interest was the birthweight of infants born to study participants; study design called for the random allocation of participants to one of three treatment groups. The authors then state: "The required size of the treatment groups was calculated from the following statistics: the standard deviation of birthweight ... is of the order of 500 g. An increment of 120 g in birthweight was arbitrarily taken to constitute a biologically meaningful gain. Given an expected difference between subjects and controls of 120 g, the required sample size for each group, in order to have a 5% risk of falsely rejecting, and a 20% risk of falsely accepting the null hypothesis, is about 320."

Table 5.6 Effect of Nifedipine on Patients with Raynaud's Phenomenon

Case	Age (yr)/ Gender	Diagnosis[a]	History of Digital Ulcer	ANA[b]	Duration of Raynaud's Phenomenon (yr)	Placebo		Nifedipine	
						Total Number of Attacks in 2 Weeks	Patient Assessment of Therapy[c]	Total Number of Attacks in 2 Weeks	Patient Assessment of Therapy[c]
1	49/F	R[d]	No	20	4	15	0	0	3+
2	20/F	R	No	Neg	3	3	1+	5	0
3	23/F	R	No	Neg	8	14	2+	6	2+
4	33/F	R	No	640	5	6	0	0	3+
5	31/F	R[d]	No	2560	2	12	0	2	3+
6	52/F	PSS	No	320	3	6	1+	1	0
7	45/M	PSS[d]	Yes	320	4	3	1+	2	2+
8	49/F	PSS	Yes	320	4	22	0	30	1+
9	29/M	PSS	Yes	1280	7	15	0	14	1+
10	33/F	PSS[d]	No	2560	9	11	1+	5	1+
11	36/F	PSS	Yes	2560	13	7	2+	2	3+
12	33/F	PSS[d]	Yes	2560	11	12	0	4	2+
13	39/F	PSS	No	320	6	45	0	45	0
14	39/M	PSS	Yes	80	6	14	1+	15	2+
15	32/F	SLE[d]	Yes	1280	5	35	1+	31	2+

Source: Data from Rodeheffer et al. [1983].

[a]R Raynaud's phenomenon without systemic disease; PSS, Raynaud's phenomenon with progressive systemic sclerosis; SLE, Raynaud's phenomenon with systemic lupus erythematosus (in addition, this patient had cryoglobulinemia).

[b]Reciprocal of antinuclear antibody titers.

[c]The Wilcoxon signed rank test, two-tailed, was performed on the patient assessment of placebo vs. nifedipine therapy: $p = 0.02$. Global assessment scale: 1− = worse; 0 = no change; 1+ = minimal improvement; 2+ = moderate improvement; and 3+ = marked improvement.

[d]Previous unsuccessful treatment with prazosin.

(a) What are the values for α and β?

(b) What is the estimate of Δ, the standardized difference?

(c) The wording in the paper suggests that sample size calculations are based on a two-sample test. Is the test one-tailed or two?

(d) Using a one-tailed test, verify that the sample size per group is $n = 215$. The number 320 reflects adjustments for losses and, perhaps, "multiple comparisons" since there are three groups (see Chapter 12).

5.16 This problem deals with the data of Problem 5.14. In column 4 of Table 5.6, patients are divided into those with a history of digital ulcers and those without. We want to compare these two groups. There are seven patients with a history and eight without.

(a) Consider the total number of attacks (in column 9) on the active drug. Carry out a two-sample t-test. Compare the group with a digital ulcer history with the group without this history. State your assumptions and conclusions.

(b) Rank all the observations in column 9, then separate the ranks into the two groups defined in part (a). Now carry out a two-sample t-test on the ranks. Compare your conclusions with those of part (b). Name an advantage to this approach. Name a disadvantage to this approach.

(c) We now do the following: Take the difference between the "placebo" and "nifedipine" columns and repeat the procedures of parts (a) and (b). Supposing that the conclusions of part (a) are not the same as those in this part, how would you interpret such discrepancies?

(d) The test carried out in part (c) is often called a *test for interaction*. Why do you suppose that this is so?

REFERENCES

Bednarek, F. J., and Roloff, D. W. [1976]. Treatment of apnea of prematurity with aminophylline. *Pediatrics*, **58**: 335–339. Used with permission.

Chen, J. R., Francisco, R. B., and Miller, T. E. [1977]. Legionnaires' disease: nickel levels. *Science*, **196**: 906–908. Copyright © 1977 by the AAAS.

Conover, W. J., and Iman, R. L. [1981]. Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, **35**: 124–129.

Dobson, J. C., Kushida, E., Williamson, M., and Friedman, E. G. [1976]. Intellectual performance of 36 phenylketonuria patients and their non-affected siblings. *Pediatrics*, **58**: 53–58. Used with permission.

Holtzman, N. A., Welcher, D. M., and Mellits, E. D. [1975]. Termination of restricted diet in children with phenylketonuria: a randomized controlled study. *New England Journal of Medicine*, **293**: 1121–1124.

Kapitulnik, J., Levin, W., Poppers, J., Tomaszewski, J. E., Jerina, D. M., and Conney, A. H. [1976]. Comparison of the hydroxylation of zoxazolamine and benzo[a]pyrene in human placenta: effect of cigarette smoking. *Clinical Pharmaceuticals and Therapeutics*, **20**: 557–564.

Mazze, R. I., Shue, G. L., and Jackson, S. H. [1971]. Renal dysfunction associated with methoxyflurane anesthesia. *Journal of the American Medical Association*, **216**: 278–288. Copyright © 1971 by the American Medical Association.

Rickman, R., Mitchell, N., Dingman, J., and Dalen, J. E. [1974]. Changes in serum cholesterol during the Stillman diet. *Journal of the American Medical Association*, **228**: 54–58. Copyright © 1974 by the American Medical Association.

Rodeheffer, R. J., Romner, J. A., Wigley, F., and Smith, C. R. [1983]. Controlled double-blind trial of Nifedipine in the treatment of Raynaud's phenomenon. *New England Journal of Medicine*, **308**: 880–883.

Rush, D., Stein, Z., and Susser, M. [1973]. The rationale for, and design of, a randomized controlled trial of nutritional supplementation in pregnancy. *Nutritional Reports International*, **7**: 547–553. Used with permission of the publisher, Butterworth-Heinemann.

Schechter, P. J., Horwitz, D., and Henkin, R. I. [1973]. Sodium chloride preference in essential hypertension. *Journal of the American Medical Association*, **225**: 1311–1315. Copyright © 1973 by The American Medical Association.

Squires, K. C., Chen, N. S., and Starr, A. [1978]. Acute effects of alcohol on auditory brainstem potentials in humans. *Science*, **201**: 174–176.

Thompson, G. L. [1991]. A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association*, **86**: 410–419.

Zelazo, P. R., Zelazo, N. A., and Kolb, S. [1972]. "Walking" in the newborn. *Science*, **176**: 314–315.

CHAPTER 6

Counting Data

6.1 INTRODUCTION

From previous chapters, recall the basic ideas of statistics. *Descriptive statistics* present data, usually in summary form. Appropriate *models* describe data concisely. The model *parameters* are *estimated* from the data. *Standard errors* and *confidence intervals* quantify the precision of estimates. Scientific hypotheses may be tested. A *formal hypothesis test* involves four things: (1) planning an experiment, or recognizing an opportunity, to collect appropriate data; (2) selecting a *significance level* and *critical region*; (3) collecting the data; and (4) rejecting the *null hypothesis* being tested if the value of the test statistic falls into the critical region. A less formal approach is to compute the *p-value*, a measure of how plausibly the data agree with the null hypothesis under study. The remainder of this book shows you how to apply these concepts in different situations, starting with the most basic of all data: counts.

Throughout recorded history people have been able to count. The word *statistics* comes from the Latin word for "state"; early statistics were counts used for the purposes of the state. Censuses were conducted for military and taxation purposes. Modern statistics is often dated from the 1662 comments on the Bills of Mortality in London. The Bills of Mortality counted the number of deaths due to each cause. John Graunt [1662] noticed patterns of regularity in the Bills of Mortality (see Section 3.3.1). Such vital statistics are important today for assessing the public health. In this chapter we return to the origin of statistics by dealing with data that arise by counting the number of occurrences of some event.

Count data lead to many different models. The following sections present examples of count data. The different types of count data will each be presented in three steps. First, you learn to recognize count data that fit a particular model. (This is the diagnosis phase.) Second, you examine the model to be used. (You learn about the illness.) Third, you learn the methods of analyzing data using the model. (At this stage you learn how to treat the disease.)

6.2 BINOMIAL RANDOM VARIABLES

6.2.1 Recognizing Binomial Random Variables

Four conditions characterize binomial data:

1. A response or trait takes on one and only one of two possibilities. Such a response is called a *binary response*. Examples are:

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

a. In a survey of the health system, people are asked whether or not they have hospitalization insurance.

b. Blood samples are tested for the presence or absence of an antigen.

c. Rats fed a potential carcinogen are examined for tumors.

d. People are classified as having or not having cleft lip.

e. Injection of a compound does or does not cause cardiac arrhythmia in dogs.

f. Newborn children are classified as having or not having Down syndrome.

2. The response is observed a known number of times. Each observation of the response is sometimes called a *Bernoulli trial*. In condition 1(a) the number of trials is the number of people questioned. In 1(b), each blood sample is a trial. Each newborn child constitutes a trial in 1(f).

3. The chance, or probability, that a particular outcome occurs is the same for each trial. In a survey such as 1(a), people are sampled at random from the population. Since each person has the same chance of being interviewed, the probability that the person has hospitalization insurance is the same in each case. In a laboratory receiving blood samples, the samples could be considered to have the same probability of having an antigen *if* the samples arise from members of a population who submit tests when "randomly" seeking medical care. The samples would not have the same probability if batches of samples arrive from different environments: for example, from schoolchildren, a military base, and a retirement home.

4. The outcome of one trial must not be influenced by the outcome of other trials. Using the terminology of Chapter 5, the trials outcomes are independent random variables. In 1(b), the trials would not be independent if there was contamination between separate blood samples. The newborn children of 1(f) might be considered independent trials for the occurrence of Down syndrome if each child has different parents. If multiple births are in the data set, the assumption of independence would not be appropriate.

We illustrate and reinforce these ideas by examples that may be modeled by the binomial distribution.

Example 6.1. Weber et al. [1976] studied the irritating effects of cigarette smoke. Sixty subjects sat, in groups of five to six, in a 30-m^2 climatic chamber. Tobacco smoke was produced by a smoking machine. After 10 cigarettes had been smoked, 47 of the 60 subjects reported that they wished to leave the room.

Let us consider the appropriateness of the binomial model for these data. Condition 1 is satisfied. Each subject was to report whether or not he or she desired to leave the room. The answer gives one of two possibilities: yes or no. Sixty trials are observed to take place (i.e., condition 2 is satisfied).

The third condition requires that each subject have the same probability of "wishing to leave the room." The paper does not explain how the subjects were selected. Perhaps the authors advertised for volunteers. In this case, the subjects might be considered "representative" of a larger population who would volunteer. The probability would be the *unknown* probability that a person selected at random from this larger population would wish to leave the room.

As we will see below, the binomial model is often used to make inferences about the unknown probability of an outcome in the "true population." Many would say that an experiment such as this shows that cigarette smoke irritates people. The extension from the ill-defined population of this experiment to humankind in general does *not* rest on this experiment. It must be based on other formal or informal evidence that humans do have much in common; in particular, one would need to assume that if one portion of humankind is irritated by cigarette smoke, so will other segments. Do you think such inferences are reasonable?

The fourth condition needed is that the trials are independent variables. The authors report in detail that the room was cleared of all smoke between uses of the climatic chamber. There should not be a carryover effect here. Recall that subjects were tested in groups of five or six. How do you think that one person's response would be changed if another person were coughing? Rubbing the eyes? Complaining? It seems possible that condition 4 is not fulfilled; that is, it seems possible that the responses were not independent.

In summary, a binomial model might be used for these data, but with some reservation. The overwhelming majority of data collected on human populations is collected under less than ideal conditions; a subjective evaluation of the worth of an experiment often enters in.

Example 6.2. Karlowski et al. [1975] reported on a controlled clinical trial of the use of ascorbic acid (vitamin C) for the common cold. Placebo and vitamin C were randomly assigned to the subjects; the experiment was to be a double-blind trial. It turned out that some subjects were testing their capsules and claimed to know the medication. Of 64 subjects who tested the capsule and guessed at the treatment, 55 were correct. Could such a split arise by chance if testing did not help one to guess correctly?

One thinks of using a binomial model for these data since there is a binary response (correct or incorrect guess) observed on a known number of people. Assuming that people tested only their own capsules, the guesses should be statistically independent. Finally, if the guesses are "at random," each subject should have the same probability—one-half—of making a correct guess since half the participants receive vitamin C and half a placebo. This binomial model would lead to a test of the hypothesis that the probability of a correct guess was 1/2.

Example 6.3. Bucher et al. [1976] studied the occurrence of hemolytic disease in newborns resulting from ABO incompatibility between the parents. Parents are said to be incompatible if the father has antigens that the mother lacks. This provides the opportunity for production of maternal antibodies from fetal–maternal stimulation. Low-weight immune antibodies that cross the placental barrier apparently cause the disease [Cavalli-Sforza and Bodmer, 1999]. The authors reviewed 7464 consecutive infants born at North Carolina Hospital. Of 3584 "black births," 43 had ABO hemolytic disease. What can be said about the true probability that a black birth has ABO hemolytic disease?

It seems reasonable to consider the number of ABO hemolytic disease cases to be binomial. The presence of disease among the 3584 trials should be independent (assuming that no parents had more than one birth during the period of case recruitment—October 1965 to March 1973—and little or no effect from kinship of parents). The births may conceptually be thought of as a sample of the population of "potential" black births during the given time period at the hospital.

6.2.2 Binomial Model

In speaking about a Bernoulli trial, without reference to a particular example, it is customary to label one outcome as a "success" and the other outcome as a "failure." The mathematical model for the binomial distribution depends on two parameters: n, the number of trials, and π, the probability of a success in one trial. A binomial random variable, say Y, is the count of the number of successes in the n trials. Of course, Y can only take on the values $0, 1, 2, \ldots, n$. If π, the probability of a success, is large (close to 1), then Y, the number of successes, will tend to be large. Conversely, if the probability of success is small (near zero), Y will tend to be small.

To do statistical analysis with binomial variables, we need the probability distribution of Y. Let k be an integer between 0 and n inclusive. We need to know $P[Y = k]$. In other words, we want the probability of k successes in n independent trials when π is the probability of success. The symbol $b(k; n, \pi)$ will be used to denote this probability. The answer involves the *binomial coefficient*. The binomial coefficient $\binom{n}{k}$ is the number of different ways that

k objects may be selected from n objects. (Problem 6.24 helps you to derive the value of $\binom{n}{k}$.) For each positive integer n, n factorial (written $n!$) is defined to be $1 \times 2 \times \cdots \times n$. So $6! = 1 \times 2 \times 3 \times 4 \times 5 \times 6 = 720$. 0!, zero factorial, is defined to be 1. With this notation the binomial coefficient may be written

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!} = \frac{n(n-1)\cdots(k+1)}{(n-k)(n-k-1)\cdots 1} \tag{1}$$

Example 6.4. This is illustrated with the following two cases:

1. Of 10 residents, three are to be chosen to cover a hospital service on a holiday. In how many ways may the residents be chosen? The answer is

$$\binom{10}{3} = \frac{10!}{7!\,3!} = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8 \times 9 \times 10}{(1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7)(1 \times 2 \times 3)} = 120$$

2. Of eight consecutive patients, four are to be assigned to drug A and four to drug B. In how many ways may the assignments be made? Think of the eight positions as eight objects; we need to choose four for the drug A patients. The answer is

$$\binom{8}{4} = \frac{8!}{4!\,4!} = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8}{(1 \times 2 \times 3 \times 4)(1 \times 2 \times 3 \times 4)} = 70$$

The binomial probability, $b(k; n, \pi)$, may be written

$$b(k; n, \pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k} \tag{2}$$

Example 6.5. Ten patients are treated surgically. For each person there is a 70% chance of successful surgery (i.e., $\pi = 0.7$). What is the probability of only five or fewer successful surgeries?

$$\begin{aligned} P[\text{five or fewer successful cases}] &= P[\text{five successful cases}] + P[\text{four successful cases}] \\ &\quad + P[\text{three successful cases}] + P[\text{two successful cases}] \\ &\quad + P[\text{one successful case}] + P[\text{no successful case}] \\ &= b(5; 10, 0.7) + b(4; 10, 0.7) + b(3; 10, 0.7) + b(2; 10, 0.7) \\ &\quad + b(1; 10, 0.7) + b(0; 10, 0.7) \\ &= 0.1029 + 0.0368 + 0.0090 + 0.0014 + 0.0001 + 0.0000 \\ &= 0.1502 \end{aligned}$$

(*Note*: The actual value is 0.1503; the answer 0.1502 is due to round-off error.)

The binomial probabilities may be calculated directly or found by a computer program. The mean and variance of a binomial random variable with parameters π and n are given by

$$\begin{aligned} E(Y) &= n\pi \\ \text{var}(Y) &= n\pi(1-\pi) \end{aligned} \tag{3}$$

From equation (3) it follows that Y/n has the expected value π:

$$E\left(\frac{Y}{n}\right) = \pi \tag{4}$$

In other words, the proportion of successes in n binomial trials is an unbiased estimate of the probability of success.

6.2.3 Hypothesis Testing for Binomial Variables

The hypothesis-testing framework established in Chapter 4 may be used for the binomial distribution. There is one minor complication. The binomial random variable can take on only a finite number of values. Because of this, it may not be possible to find hypothesis tests such that the significance level is precisely some fixed value. If this is the case, we construct regions so that the significance level is close to the desired significance level.

In most situations involving the binomial distribution, the number of trials (n) is known. We consider statistical tests about the true value of π. Let $p = Y/n$. If π is hypothesized to be π_0, an observed value of p close to π_0 reinforces the hypothesis; a value of p differing greatly from π_0 makes the hypothesis seem unlikely.

Procedure 1. To construct a significance test of H_0: $\pi = \pi_0$ against H_A: $\pi \neq \pi_0$, at significance level α:

1. Find the smallest c such that $P\ [|p - \pi_0| \geq c] \leq \alpha$ when H_0 is true.
2. Compute the *actual* significance level of the test; the actual significance level is $P[|p - \pi_0| \geq c]$.
3. Observe p, call it $\widehat{p}$; reject H_0 if $|\widehat{p} - \pi_0| \geq c$.

The quantity c is used to determine the critical value (see Definition 4.19); that is, determine the bounds of the rejection region, which will be $\pi_0 \pm c$. Equivalently, working in the Y scale, the region is defined by $n\pi_0 \pm nc$.

Example 6.6. For $n = 10$, we want to construct a test of the null hypothesis H_0: $\pi = 0.4$ vs. the alternative hypothesis H_A: $\pi \neq 0.4$. Thus, we want a two-sided test. The significance level is to be as close to $\alpha = 0.05$ as possible. We work in the $Y = np$ scale. Under H_0, Y has mean $n\pi = (10)(0.4) = 4$. We want to find a value C such that $P[|Y - 4| \geq C]$ is as close to $\alpha = 0.05$ (and less than α) as possible. The quantity C is the distance Y is from the null hypothesis value 4. Using the definition of the binomial distribution, we construct Table 6.1.

The closest α-value to 0.05 is $\alpha = 0.0183$; the next value is 0.1012. Hence we choose $C = 4$; we reject the null hypothesis H_0: $n\pi = 4$ if $Y = 0$ or $Y \geq 8$; equivalently, if $p = 0$ or $p \geq 0.8$, or in the original formulation, if $|p - 0.4| \geq 0.4$ since $C = 10c$.

Table 6.1 C-Values for Example 6.6

C	$4 - C$	$C + 4$	$P[\|Y - 4\| \geq C] = \alpha$
6	—	10	$0.0001 = P[Y = 10]$
5	—	9	$0.0017 = P[Y \geq 9]$
4	0	8	$0.0183 = P[Y = 0] + P[Y \geq 8]$
3	1	7	$0.1012 = P[Y \leq 1] + P[Y \geq 7]$
2	2	6	$0.3335 = P[Y \leq 2] + P[Y \geq 6]$
1	3	5	$0.7492 = P[Y \leq 3] + P[Y \geq 5]$

Procedure 2. To find the *p-value* for testing the hypothesis H_0: $\pi = \pi_0$ vs. H_A: $\pi \neq \pi_0$:

1. Observe $p : \widehat{p}$ is now fixed, where $\widehat{p} = y/n$.
2. Let $\widetilde{p}$ be a binomial random variable with parameters n and π_0. The p-value is $P[|\widetilde{p} - \pi_0| \geq |\widehat{p} - \pi_0|]$.

Example 6.7. Find the p-value for testing $\pi = 0.5$ if $n = 10$ and we observe that $p = 0.2$. $|\widetilde{p} - 0.5| \geq |0.2 - 0.5| = 0.3$ only if $\widetilde{p} = 0.0, 0.1, 0.2, 0.8, 0.9$, or 1.0. The p-value can be computed by software or by adding up the probabilities of the "more extreme" values: $0.0010 + 0.0098 + 0.0439 + 0.0439 + 0.0098 + 0.0010 = 0.1094$. Tables for this calculation are provided in the Web appendix. The appropriate one-sided hypothesis test and calculation of a one-sided p-value is given in Problem 6.25.

6.2.4 Confidence Intervals

Confidence intervals for a binomial proportion can be found by computer or by looking up the confidence limits in a table. Such tables are not included in this book, but are available in any standard handbook of statistical tables, for example, Odeh et al. [1977], Owen [1962], and Beyer [1968].

6.2.5 Large-Sample Hypothesis Testing

The central limit theorem holds for binomial random variables. If Y is binomial with parameters n and π, then for "large n,"

$$\frac{Y - E(Y)}{\sqrt{\text{var}(Y)}} = \frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}}$$

has approximately the same probability distribution as an $N(0, 1)$ random variable. Equivalently, since $Y = np$, the quantity $(p - \pi)/\sqrt{\pi(1-\pi)/n}$ approaches a normal distribution. We will work interchangeably in the p scale or the Y scale. For large n, hypothesis tests and confidence intervals may be formed by using critical values of the standard normal distribution.

The closer π is to 1/2, the better the normal approximation will be. If $n \leq 50$, it is preferable to use tables for the binomial distribution and hypothesis tests as outlined above. A reasonable rule of thumb is that n is "large" if $n\pi(1-\pi) \geq 10$.

In using the central limit theorem, we are approximating the distribution of a discrete random variable by the continuous normal distribution. The approximation can be improved by using a *continuity correction*. The normal random variable with continuity correction is given by

$$Z_c = \begin{cases} \dfrac{Y - n\pi - 1/2}{\sqrt{n\pi(1-\pi)}} & \text{if } Y - n\pi > 1/2 \\ \dfrac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} & \text{if } |Y - n\pi| \leq 1/2 \\ \dfrac{Y - n\pi + 1/2}{\sqrt{n\pi(1-\pi)}} & \text{if } Y - n\pi < -1/2 \end{cases}$$

For $n\pi(1-\pi) \geq 100$, or quite large, the factor of 1/2 is usually ignored.

Procedure 3. Let Y be binomial n, π, with a large n. A hypothesis test of H_0: $\pi = \pi_0$ vs. H_A: $\pi \neq \pi_0$ at significance level α is given by computing Z_c with $\pi = \pi_0$. The null hypothesis is rejected if $|Z_c| \geq z_{1-\alpha/2}$.

Example 6.8. In Example 6.2, of the 64 persons who tested their capsules, 55 guessed the treatment correctly. Could so many people have guessed the correct treatment "by chance"? In Example 6.2 we saw that chance guessing would correspond to $\pi_0 = 1/2$. At a 5% significance level, is it plausible that $\pi_0 = 1/2$?

As $n\pi_0(1 - \pi_0) = 64 \times 1/2 \times 1/2 = 16$, a large-sample approximation is reasonable. $y - n\pi_0 = 55 - 64 \times 1/2 = 23$, so that

$$Z_c = \frac{Y - n\pi_0 - 1/2}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{22.5}{\sqrt{64 \times 1/2 \times 1/2}} = 5.625$$

As $|Z_c| = 5.625 > 1.96 = z_{0.975}$, the null hypothesis that the correct guessing occurs purely by chance must be rejected.

Procedure 4. The large-sample two-sided p-value for testing H_0: $\pi = \pi_0$ vs. H_A: $\pi \neq \pi_0$ is given by $2(1 - \Phi(|Z_c|))$. $\Phi(x)$ is the probability that an $N(0, 1)$ random variable is less than x. $|Z_c|$ is the absolute value of Z_c.

6.2.6 Large-Sample Confidence Intervals

Procedure 5. For large n, say $n\widehat{p}(l-\widehat{p}) \geq 10$, an approximate $100(1-\alpha)\%$ confidence interval for π is given by

$$\left(\widehat{p} - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \quad \widehat{p} + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) \tag{5}$$

where $\widehat{p} = y/n$ is the observed proportion of successes.

Example 6.9. Find a 95% confidence interval for the true fraction of black children having ABO hemolytic disease in the population represented by the data of Example 6.3. Using formula (5) the confidence interval is

$$\frac{43}{3584} \pm 1.96\sqrt{\frac{(43/3584)(1-43/3584)}{3584}} \quad \text{or} \quad (0.0084, 0.0156)$$

6.3 COMPARING TWO PROPORTIONS

Often, one is not interested in only one proportion but wants to compare two proportions. A health services researcher may want to see whether one of two races has a higher percentage of prenatal care. A clinician may wish to discover which of two drugs has a higher proportion of cures. An epidemiologist may be interested in discovering whether women on oral contraceptives have a higher incidence of thrombophlebitis than those not on oral contraceptives. In this section we consider the statistical methods appropriate for comparing two proportions.

6.3.1 Fisher's Exact Test

Data to estimate two different proportions will arise from observations on two populations. Call the two sets of observations sample 1 and sample 2. Often, the data are presented in 2×2 (verbally, "two by two") tables as follows:

	Success	Failure
Sample 1	n_{11}	n_{12}
Sample 2	n_{21}	n_{22}

The first sample has n_{11} successes in $n_{11} + n_{12}$ trials; the second sample has n_{21} successes in $n_{21} + n_{22}$ trials. Often, the null hypothesis of interest is that the probability of success in the two populations is the same. *Fisher's exact test* is a test of this hypothesis for small samples.

The test uses the row and column totals. Let $n_{1\cdot}$ denote summation over the second index; that is, $n_{1\cdot} = n_{11} + n_{12}$. Similarly define $n_{2\cdot}$, $n_{\cdot 1}$, and $n_{\cdot 2}$. Let $n_{\cdot\cdot}$ denote summation over both indices; that is, $n_{\cdot\cdot} = n_{11} + n_{12} + n_{21} + n_{22}$. Writing the table with row and column totals gives:

	Success	**Failure**	
Sample 1	n_{11}	n_{12}	$n_{1\cdot}$
Sample 2	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

Suppose that the probabilities of success in the two populations are the same. Suppose further that we are given the row and column totals but *not* n_{11}, n_{12}, n_{21}, and n_{22}. What is the probability distribution of n_{11}?

Consider the $n_{\cdot\cdot}$ trials as $n_{\cdot\cdot}$ objects; for example, $n_{1\cdot}$ purple balls and $n_{2\cdot}$ gold balls. Since each trial has the same probability of success, any subset of $n_{\cdot 1}$ trials (balls) has the same probability of being chosen as any other. Thus, the probability that n_{11} has the value k is the same as the probability that there are k purple balls among $n_{\cdot 1}$ balls chosen without replacement from an urn with $n_{1\cdot}$ purple balls and $n_{2\cdot}$ gold balls. The probability distribution of n_{11} is called the *hypergeometric distribution*.

The mathematical form of the hypergeometric probability distribution is derived in Problem 6.26.

Example 6.10. Kennedy et al. [1981] consider patients who have undergone coronary artery bypass graft surgery (CABG). CABG takes a saphenous vein from the leg and connects the vein to the aorta, where blood is pumped from the heart, and to a coronary artery, an artery that supplies the heart muscle with blood. The vein is placed beyond a narrowing, or stenosis, in the coronary artery. If the artery would close at the narrowing, the heart muscle would still receive blood. There is, however, some risk to this open heart surgery. Among patients with moderate narrowing (50 to 74%) of the left main coronary artery emergency cases have a high surgical mortality rate. The question is whether emergency cases have a surgical mortality different from that of nonemergency cases. The in-hospital mortality figures for emergency surgery and other surgery were:

	Discharge Status	
Surgical Priority	**Dead**	**Alive**
Emergency	1	19
Other	7	369

From the hypergeometric distribution, the probability of an observation this extreme is $0.3419 = P[n_{11} \geq 1] = P[n_{11} = 1] + \cdots + P[n_{11} = 8]$. (Values for $n_{\cdot\cdot}$ this large are not tabulated and need to be computed directly.) These data do not show any difference beyond that expected by chance.

Example 6.11. Sudden infant death syndrome (SIDS), or crib death, results in the unexplained death of approximately two of every 1000 infants during their first year of life. To study the genetic component of such deaths, Peterson et al. [1980] examined sets of twins with at least one SIDS child. If there is a large genetic component, the probability of both twins dying will

be larger for identical twin sets than for fraternal twin sets. If there is no genetic component, but only an environmental component, the probabilities should be the same. The following table gives the data:

	SIDS Children	
Type of Twin	**One**	**Both**
Monozygous (identical)	23	1
Dizygous (fraternal)	35	2

The Fisher's exact test one-sided p-value for testing that the probability is higher for monozygous twins is $p = 0.784$. Thus, there is no evidence for a genetic component in these data.

6.3.2 Large-Sample Tests and Confidence Intervals

As mentioned above, in many situations one wishes to compare proportions as estimated by samples from two populations to see if the true population parameters might be equal or if one is larger than the other. Examples of such situations are a drug and placebo trial comparing the percentage of patients experiencing pain relief; the percentage of rats developing tumors under diets involving different doses of a food additive; and an epidemiologic study comparing the percentage of infants suffering from malnutrition in two countries.

Suppose that the first binomial variable (the sample from the first population) is of size n_1 with probability π_1, estimated by the sample proportion p_1. The second sample estimates π_2 by p_2 from a sample of size n_2.

It is natural to compare the proportions by the difference $p_1 - p_2$. The mean and variance are given by

$$E(p_1 - p_2) = \pi_1 - \pi_2,$$

$$\text{var}(p_1 - p_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

A version of the central limit theorem shows that for large n_1 and n_2 [say, both $n_1\pi_1(1-\pi_1)$ and $n_2\pi_2(1-\pi_2)$ greater than 10],

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} = Z$$

is an approximately normal pivotal variable. From this, hypothesis tests and confidence intervals develop in the usual manner, as illustrated below.

Example 6.12. The paper by Bucher et al. [1976] discussed in Example 6.3 examines racial differences in the incidence of ABO hemolytic disease by examining records for infants born at North Carolina Memorial Hospital. In this paper a variety of possible ways of defining hemolytic disease are considered. Using their class I definition, the samples of black and white infants have the following proportions with hemolytic disease:

$$\text{black infants, } n_1 = 3584, \qquad p_1 = \frac{43}{3584}$$

$$\text{white infants, } n_2 = 3831, \qquad p_2 = \frac{17}{3831}$$

It is desired to perform a two-sided test of the hypothesis $\pi_1 = \pi_2$ at the $\alpha = 0.05$ significance level. The test statistic is

$$Z = \frac{(43/3584) - (17/3831)}{\sqrt{[(43/3584)(1 - 43/3584)]/3584 + [(17/3831)(1 - 17/3831)]/3831}} \doteq 3.58$$

The two-sided p-value is $P[|Z| \geq 3.58] = 0.0003$ from Table A.1. As $0.0003 < 0.05$, the null hypothesis of equal rates, $\pi_1 = \pi_2$, is rejected at the significance level 0.05.

The pivotal variable may also be used to construct a confidence interval for $\pi_1 - \pi_2$. Algebraic manipulation shows that the endpoints of a symmetric (about $p_1 - p_2$) confidence interval are given by

$$p_1 - p_2 \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

For a 95% confidence interval $z_{1-\alpha/2} = 1.96$ and the interval for this example is

$$0.00756 \pm 0.00414 \quad \text{or} \quad (0.00342, 0.01170)$$

A second statistic for testing for equality in two proportions is the χ^2 (chi-square) statistic. This statistic is considered in more general situations in Chapter 7. Suppose that the data are as follows:

	Sample 1	**Sample 2**	
Success	$n_1 p_1 = n_{11}$	$n_2 p_2 = n_{12}$	$n_{1\cdot}$
Failure	$n_1(1 - p_1) = n_{21}$	$n_2(1 - p_2) = n_{22}$	$n_{2\cdot}$
	$n_1 = n_{\cdot 1}$	$n_2 = n_{\cdot 2}$	$n_{\cdot\cdot}$

A statistic for testing H_0: $\pi_1 = \pi_2$ is the χ^2 statistic with one degree of freedom. It is calculated by

$$X^2 = \frac{n_{\cdot\cdot}(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$$

For technical reasons (Note 6.2) the chi-square distribution with continuity correction, designated by X_c^2, is used by some people. The formula for X_c^2 is

$$X_c^2 = \frac{n_{\cdot\cdot}(|n_{11}n_{22} - n_{12}n_{21}| - \frac{1}{2}n_{\cdot\cdot})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$$

For the Bucher et al. [1976] data, the values are as follows:

	Race		
ABO Hemolytic Disease	**Black**	**White**	**Total**
Yes	43	17	60
No	3541	3814	7355
Total	3584	3831	7415

$$X^2 = \frac{7415(43 \times 3814 - 17 \times 3541)^2}{60(7355)(3584)(3831)} = 13.19$$

$$X_c^2 = \frac{7415(|43 \times 3814 - 17 \times 3541| - 7415/2)^2}{60(7355)(3584)(3831)} = 12.26$$

These statistics, for large n, have a chi-square (χ^2) distribution with one degree of freedom under the null hypothesis of equal proportions. If the null hypothesis is not true, X^2 or X_c^2 will tend to be large. The null hypothesis is rejected for large values of X^2 or X_c^2. Table A.3 has χ^2 critical values. The Bucher data have $p < 0.001$ since the 0.001 critical value is 10.83 and require rejection of the null hypothesis of equal proportions.

From Note 5.3 we know that $\chi_1^2 = Z^2$. For this example the value of $Z^2 = 3.58^2 = 12.82$ is close to the value $X^2 = 13.19$. The two values would have been identical (except for rounding) if we had used in the calculation of Z an estimate of the standard error of $\sqrt{pq(1/n_1 + 1/n_2)}$, where $p = 60/7415$ is the pooled estimate of π under the null hypothesis $\pi_1 = \pi_2 = \pi$.

6.3.3 Finding Sample Sizes Needed for Testing the Difference between Proportions

Consider a study planned to test the equality of the proportions π_1 and π_2. Only studies in which both populations are sampled the same number of times, $n = n_1 = n_2$, will be considered here. There are five quantities that characterize the performance and design of the test:

1. π_1, the proportion in the first population.
2. π_2, the proportion in the second population under the alternative hypothesis.
3. n, the number of observations to be obtained from *each* of the two populations.
4. The significance level α at which the statistical test will be made. α is the probability of rejecting the null hypothesis when it is true. The null hypothesis is that $\pi_1 = \pi_2$.
5. The probability, β, of accepting the null hypothesis when it is not true, but the alternative is true. Here we will have $\pi_1 \neq \pi_2$ under the alternative hypothesis (π_1 and π_2 as specified in quantities 1 and 2 above).

These quantities are interrelated. It is not possible to change one of them without changing at least one of the others. The actual determination of sample size is usually an iterative process; the usual state of affairs is that the desire for precision and the practicality of obtaining an appropriate sample size are in conflict. In practice, one usually considers various possible combinations and arrives at a "reasonable" sample size or decides that it is not possible to perform an adequate experiment within the constraints involved.

The "classical" approach is to specify π_1, π_2 (for the alternative hypothesis), α, and β. These parameters determine the sample size n. Table A.8 gives some sample sizes for such binomial studies using one-sided hypothesis tests (see Problem 6.27). An approximation for n is

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

where $\alpha = 1 - \Phi(z_{1-\alpha})$; that is, $z_{1-\alpha}$ is the value such that a $N(0, 1)$ variable Z has $P[Z > z_{1-\alpha}] = \alpha$. In words, $z_{1-\alpha}$ is the one-sided normal α critical value. Similarly, $z_{1-\beta}$ is the one-sided normal β critical value.

Figure 6.1 is a flow diagram for calculating sample sizes for discrete (binomial) as well as continuous variables. It illustrates that the format is the same for both: first, values of α and β are selected. A one- or two-sided test determines $z_{1-\alpha}$ or $z_{1-\alpha/2}$ and the quantity NUM, respectively. For discrete data, the quantities π_1 and π_2 are specified, and

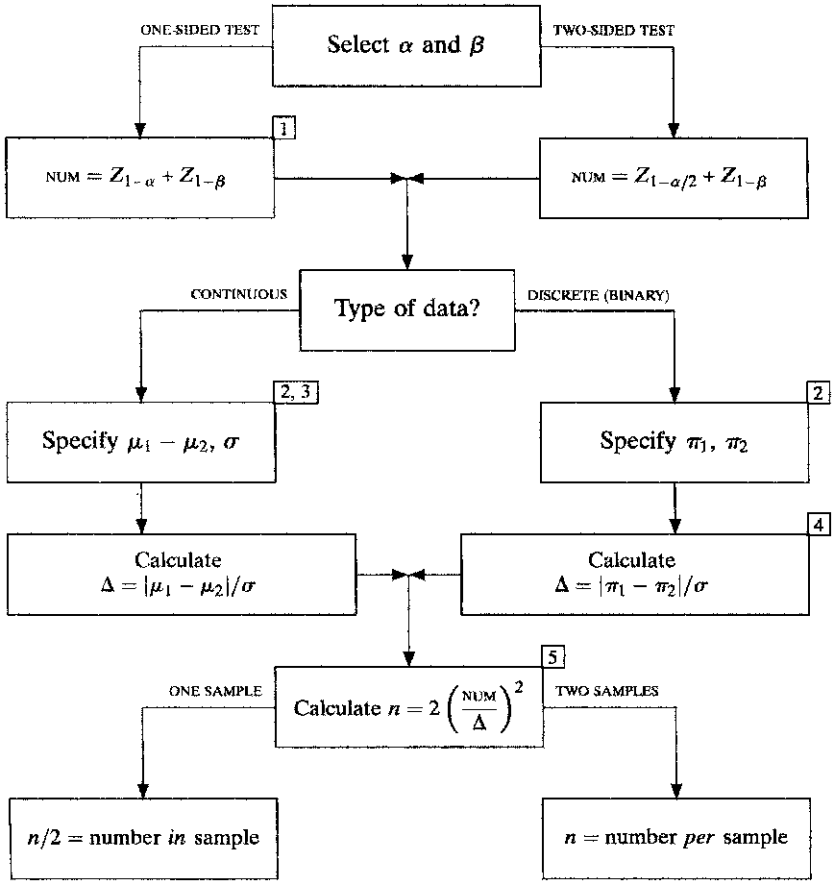


Figure 6.1 Flowchart for sample-size calculations (continuous and discrete variables).

1 Values of Z_c for various values of c are:

c	0.500	0.800	0.900	0.950	0.975	0.990	0.995
Z_c	0.000	0.842	1.282	1.645	1.960	2.326	2.576

2 If one sample, μ_2 and π_2 are null hypothesis values.

3 If $\sigma_1^2 \neq \sigma_2^2$, calculate $\sigma^2 = \frac{1}{2}(\sigma_1^2 + \sigma_2^2)$.

4 $\sigma = \sqrt{\frac{1}{2}(\pi_1(1-\pi_1) + \pi_2(1-\pi_2))}$.

5 Sample size for discrete case is an approximation. For an improved estimate, use $n^* = n + 2/\Delta$.

Note: Two sample case, unequal sample sizes. Let n_1 and kn_1 be the required sample sizes. Calculate n as before. Then calculate $n_1 = n(k+1)/2k$ and $n_2 = kn_1$. (Total sample size will be larger.) If also, $\sigma_1^2 \neq \sigma_2^2$ calculate n using σ_1; then calculate $n_1 = (n/2)[1 + \sigma_2^2/(k\sigma_1^2)]$ and $n_2 = kn_1$.

$\Delta = |\pi_1 - \pi_2|/\sqrt{\frac{1}{2}(\pi_1(1-\pi_1) + \pi_2(1-\pi_2))}$ is calculated. This corresponds to the standardized differences $\Delta = |\mu_1 - \mu_2|/\sigma$ associated with normal or continuous data. The quantity $n = 2(\text{NUM}/\Delta)^2$ then produces the sample size needed for a two-sample procedure. For a one-sample procedure, the sample size is $n/2$. Hence a two-sample procedure requires a total of *four* times the number of observations of the one-sample procedure. Various refinements are available

in Figure 6.1. A list of the most common Z-values is provided. If a one-sample test is wanted, the values of μ_2 and π_2 can be considered the null hypothesis values. Finally, the equation for the sample size in the discrete case is an approximation, and a more precise estimate, n^*, can be obtained from

$$n^* = n + \frac{2}{\Delta}$$

This formula is reasonably accurate.

Other approaches are possible. For example, one might specify the largest feasible sample size n, α, π_1, and π_2 and then determine the power $1 - \beta$. Figure 6.2 relates π_1, $\Delta = \pi_2 - \pi_1$, and n for two-sided tests for $\alpha = 0.05$ and $\beta = 0.10$.

Finally, we note that in certain situations where sample size is necessarily limited, for example, a rare cancer site with several competing therapies, trials with $\alpha = 0.10$ and $\beta = 0.50$ have been run.

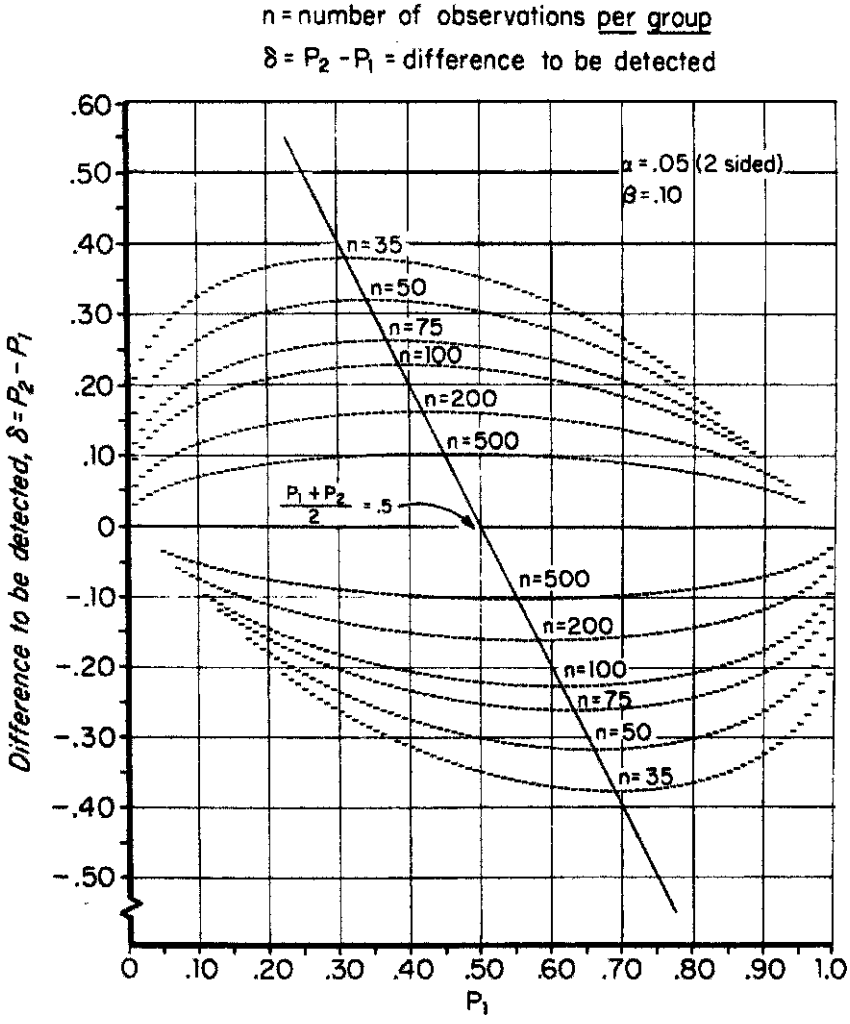


Figure 6.2 Sample sizes required for testing two proportions, π_1 and π_2 with 90% probability of obtaining a significant result at the 5% (two-sided) level. (From Feigl [1978].)

In practice, it is difficult to arrive at a sample size. To one unacquainted with statistical ideas, there is a natural tendency to expect too much from an experiment. In answer to the question, "What difference would you like to detect?" the novice will often reply, "any difference," leading to an infinite sample size!

6.3.4 Relative Risk and the Odds Ratio

In this section we consider studies looking for an association between two binary variables, that is, variables that take on only two outcomes. For definiteness we speak of the two variables as disease and exposure, since the following techniques are often used in epidemiologic and public health studies. In effect we are comparing two proportions (the proportions with disease) in two populations: those with and without exposure. In this section we consider methods of summarizing the association.

Suppose that one had a complete enumeration of the population at hand and the true proportions in the population. The probabilities may be presented in a 2×2 table:

	Disease	
Exposure	**+ (Yes)**	**− (No)**
+ (Yes)	π_{11}	π_{12}
− (No)	π_{21}	π_{22}

where $\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1$.

There are subtleties glossed over here. For example, by disease (for a human population), does one mean that the person develops the disease at some time before death, has the disease at a particular point in time, or develops it by some age? This ignores the problems of accurate diagnosis, a notoriously difficult problem. Similarly, exposure needs to be defined carefully as to time of exposure, length of exposure, and so on.

What might be a reasonable measure of the effect of exposure? A plausible comparison is $P[\text{disease}+|\text{exposure}+]$ with $P[\text{disease}+|\text{exposure}-]$. In words, it makes sense to compare the probability of having disease among those exposed with the probability of having the disease among those not exposed.

Definition 6.1. A standard measure of the strength of the exposure effect is the *relative risk*. The relative risk is defined to be

$$\rho = \frac{P[\text{disease}+|\text{exposure}+]}{P[\text{disease}+|\text{exposure}-]} = \frac{\pi_{11}/(\pi_{11}+\pi_{12})}{\pi_{21}/(\pi_{21}+\pi_{22})} = \frac{\pi_{11}(\pi_{21}+\pi_{22})}{\pi_{21}(\pi_{11}+\pi_{12})}$$

Thus, a relative risk of 5 means that an exposed person is five times as likely to have the disease. The following tables of proportions or probabilities each has a relative risk of 2:

	Disease		**Disease**		**Disease**		**Disease**	
Exposure	**+**	**−**	**+**	**−**	**+**	**−**	**+**	**−**
+	0.50	0.00	0.25	0.25	0.10	0.40	0.00010	0.49990
−	0.25	0.25	0.125	0.375	0.05	0.45	0.0005	0.49995

We see that many patterns may give rise to the same relative risk. This is not surprising, as one number is being used to summarize four numbers. In particular, information on the amount of disease and/or exposure is missing.

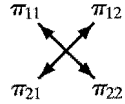
Definition 6.2. Given that one has the exposure, the *odds* (or betting odds) of getting the disease are

$$\frac{P[\text{disease} + \mid \text{exposure}+]}{P[\text{disease} - \mid \text{exposure}+]}$$

Similarly, one may define the odds of getting the disease given no exposure. Another measure of the amount of association between the disease and exposure is the *odds ratio* defined to be

$$\begin{aligned}\omega &= \frac{P[\text{disease} + \mid \text{exposure}+]/P[\text{disease} - \mid \text{exposure}+]}{P[\text{disease} + \mid \text{exposure}-]/P[\text{disease} - \mid \text{exposure}-]} \\ &= \frac{(\pi_{11}/(\pi_{11}+\pi_{12}))/(\pi_{12}/(\pi_{11}+\pi_{12}))}{(\pi_{21}/(\pi_{21}+\pi_{22}))/(\pi_{22}/(\pi_{21}+\pi_{22}))} \\ &= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\end{aligned}$$

The odds ratio is also called the *cross-product ratio* on occasion; this name is suggested by the following scheme:



Consider now how much the relative risk and odds ratio may differ by looking at the ratio of the two terms, ρ and ω,

$$\frac{\rho}{\omega} = \left(\frac{\pi_{21}+\pi_{22}}{\pi_{22}}\right)\left(\frac{\pi_{12}}{\pi_{11}+\pi_{12}}\right)$$

Suppose that the disease affects a small segment of the population. Then π_{11} is small compared to π_{12}, so that $\pi_{12}/(\pi_{11}+\pi_{12})$ is approximately equal to 1. Also, π_{21} will be small compared to π_{22}, so that $(\pi_{21}+\pi_{22})/\pi_{22}$ is approximately 1. Thus, in this case, $\rho/\omega = 1$. Restating this: If the disease affects a small fraction of the population (in both exposed and unexposed groups), the odds ratio and the relative risk are approximately equal. For this reason the odds ratio is often called the *approximate relative risk*. If the disease affects less than 5% in each group, the two quantities can be considered approximately equal.

The data for looking at the relative risk or the odds ratio usually arise in one of three ways, each of which is illustrated below. The numbers observed in each of the four cells will be denoted as follows:

	Disease	
Exposure	**+**	**−**
+	n_{11}	n_{12}
−	n_{21}	n_{22}

As before, a dot will replace a subscript when the entries for that subscript are summed. For example,

$$\begin{aligned}n_{1\cdot} &= n_{11} + n_{12} \\ n_{\cdot 2} &= n_{12} + n_{22} \\ n_{\cdot\cdot} &= n_{11} + n_{12} + n_{21} + n_{22}\end{aligned}$$

Pattern 1. (Cross-Sectional Studies: Prospective Studies of a Sample of the Population) There is a sample of size $n..$ from the population; both traits (exposure and disease) are measured on each subject. This is called *cross-sectional data* when the status of the two traits is measured at some fixed cross section in time. In this case the expected number in each cell is the expectation:

$n..\pi_{11}$	$n..\pi_{12}$	
$n..\pi_{21}$	$n..\pi_{22}$	
		$n..$

Example 6.13. The following data are from Meyer et al. [1976]. This study collected information on all births in 10 Ontario (Canada) teaching hospitals during 1960–1961. A total of 51,490 births was involved, including fetal deaths and neonatal deaths (perinatal mortality). The paper considers the association of perinatal events and maternal smoking during pregnancy. Data relating perinatal mortality and smoking are as follows:

	Perinatal Mortality		
Maternal Smoking	**Yes**	**No**	**Total**
Yes	619	20,443	21,062
No	634	26,682	27,316
Total	1,253	47,125	48,378

Estimation of the relative risk and odds ratio is discussed below.

Pattern 2. (Prospective Study: Groups Based on Exposure) In a prospective study of exposure, fixed numbers—say $n_1.$ and $n_2.$—of people with and without the exposure are followed. The endpoints are then noted. In this case the expected number of observations in the cells are:

$n_1 . \dfrac{\pi_{11}}{\pi_{11}+\pi_{12}}$	$n_1 . \dfrac{\pi_{12}}{\pi_{11}+\pi_{12}}$	$n_1.$
$n_2 . \dfrac{\pi_{21}}{\pi_{21}+\pi_{22}}$	$n_2 . \dfrac{\pi_{22}}{\pi_{21}+\pi_{22}}$	$n_2.$

Note that as the sample sizes of the exposure and nonexposure groups are determined by the experimenter, the data will not allow estimates of the proportion exposed, only the conditional probability of disease given exposure or nonexposure.

Example 6.14. As an example, consider a paper by Shapiro et al. [1974] in which they state that "by the end of this [five-year] period, there were 40 deaths in the [screened] study group of about 31,000 women as compared with 63 such deaths in a comparable group of women." Placing this in a 2×2 table and considering the screening to be the exposure, the data are:

	Breast Cancer Death		
On Study (Screened)	**Yes**	**No**	**Total**
Yes	40	30,960	31,000
No	63	30,937	31,000

Pattern 3. (Retrospective Studies) The third way of commonly collecting the data is the retrospective study. Usually, cases and an appropriate control group are identified. (Matched or paired data are *not* being discussed here.) In this case, the sizes of the disease and control groups,

$n_{\cdot 1}$ and $n_{\cdot 2}$, are specified. From such data one cannot estimate the probability of disease but rather, the probability of being exposed given that a person has the disease and the probability of exposure given that a person does not have the disease. The expected number of observations in each cell is

$$\begin{array}{cc} n_{\cdot 1}\dfrac{\pi_{11}}{\pi_{11}+\pi_{21}} & n_{\cdot 2}\dfrac{\pi_{12}}{\pi_{12}+\pi_{22}} \\ n_{\cdot 1}\dfrac{\pi_{21}}{\pi_{11}+\pi_{21}} & n_{\cdot 2}\dfrac{\pi_{22}}{\pi_{12}+\pi_{22}} \\ \hline n_{\cdot 1} & n_{\cdot 2} \end{array}$$

Example 6.15. Kelsey and Hardy [1975] studied the driving of motor vehicles as a risk factor for acute herniated lumbar intervertebral disk. Their cases were people between the ages of 20 and 64; the studies were conducted in the New Haven metropolitan area at three hospitals or in the office of two private radiologists. The cases had low-back x-rays and were interviewed and given a few simple diagnostic tests. A control group was composed of those with low-back x-rays who were not classified as surgical probable or possible cases of herniated disk and who had not had their symptoms for more than one year. The in-patients, cases, and controls, of the Yale–New Haven hospital were asked if their job involved driving a motor vehicle. The data were:

	Herniated Disk?	
Motor Vehicle Job?	**Yes (Cases)**	**No (Controls)**
Yes	8	1
No	47	26
Total	55	27

Consider a two-way layout of disease and exposure to an agent thought to influence the disease:

	Disease	
Exposure	**+**	**−**
+	n_{11}	n_{12}
−	n_{21}	n_{22}

The three types of studies discussed above can be thought of as involving conditions on the marginal totals indicated in Table 6.2.

Table 6.2 Characterization of Cross-Sectional, Prospective, and Retrospective Studies and Relationship to Possible Estimation of Relative Risk and Odds Ratio

	Totals for:		Can One Estimate the:	
Type of Study	Column	Row	Relative Risk?	Odds Ratio?
Cross-sectional or prospective sample	Random	Random	Yes	Yes
Prospective on exposure	Random	Fixed	Yes	Yes
Retrospective	Fixed	Random	No	Yes

For example, a prospective study can be thought of as a situation where the totals for "exposure+" and "exposure−" are fixed by the experimenter, and the column totals will vary randomly depending on the association between the disease and the exposure.

For each of these three types of table, how might one estimate the relative risk and/or the odds ratio? From our tables of expected numbers of observations, it is seen that for tables of types 1 and 2,

$$\frac{E(n_{11})/(E(n_{11})+E(n_{12}))}{E(n_{21})/(E(n_{21})+E(n_{22}))} = \frac{E(n_{11})/E(n_{1\cdot})}{E(n_{21})/E(n_{2\cdot})} = \frac{\pi_{11}/(\pi_{11}+\pi_{12})}{\pi_{21}/(\pi_{21}+\pi_{22})} = \rho$$

Thus, one estimates the relative risk ρ by replacing the expected value of n_{11} by the observed value of n_{11}, etc., giving

$$\widehat{\rho} = \frac{n_{11}/n_{1\cdot}}{n_{21}/n_{2\cdot}}$$

For retrospective studies of type 3 it is not possible to estimate ρ unless the disease is rare, in which case the estimate of the odds ratio gives a reasonable estimate of the relative risk.

For all three types of tables, one sees that

$$\frac{E(n_{11})E(n_{22})}{E(n_{12})E(n_{21})} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \omega$$

Therefore, we estimate the odds ratio by

$$\widehat{\omega} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

It is clear from the definition of relative risk that if exposure has no association with the disease, $\rho = 1$. That is, both "exposed" and "nonexposed" have the same probability of disease. We verify this mathematically, and also that under the null hypothesis of no association, the odds ratio ω is also 1. Under H_0:

$$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \qquad \text{for} \quad i = 1,2 \quad \text{and} \quad j = 1,2$$

Thus,

$$\rho = \frac{\pi_{11}/\pi_{1\cdot}}{\pi_{21}/\pi_{2\cdot}} = \frac{\pi_{1\cdot}\pi_{\cdot 1}/\pi_{1\cdot}}{\pi_{2\cdot}\pi_{\cdot 1}/\pi_{2\cdot}} = 1 \quad \text{and} \quad \omega = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\pi_{1\cdot}\pi_{\cdot 1}\pi_{2\cdot}\pi_{\cdot 2}}{\pi_{1\cdot}\pi_{\cdot 2}\pi_{2\cdot}\pi_{\cdot 1}} = 1$$

If ρ or ω are greater than 1, the exposed group has an increased risk of the disease. If ρ or ω are less than 1, the group not exposed has an increased risk of the disease. Note that an increased or decreased risk may, or may not, be due to a causal mechanism.

For the three examples above, let us calculate the estimated relative risk and odds ratio where appropriate. For the smoking and perinatal mortality data,

$$\widehat{\rho} = \frac{619/21{,}062}{634/27{,}316} \doteq 1.27, \qquad \widehat{\omega} = \frac{619(26{,}682)}{634(20{,}443)} \doteq 1.27$$

From these data we estimate that smoking during pregnancy is associated with an increased risk of perinatal mortality that is 1.27 times as large. (*Note*: We have not concluded that smoking causes the mortality, only that there is an association.)

The data relating screening for early detection of breast cancer and five-year breast cancer mortality gives estimates

$$\widehat{\rho} = \frac{40/31{,}000}{63/31{,}000} \doteq 0.63, \qquad \widehat{\omega} = \frac{40(30{,}937)}{63(30{,}960)} \doteq 0.63$$

Thus, in this study, screening was associated with a risk of dying of breast cancer within five years only 0.63 times as great as the risk among those not screened.

In the unmatched case–control study, only ω can be estimated:

$$\widehat{\omega} = \frac{8 \times 26}{1 \times 47} \doteq 4.43$$

It is estimated that driving jobs increase the risk of a herniated lumbar intervertebral disk by a factor of 4.43.

Might there really be no association in the tables above and the estimated $\widehat{\rho}$'s and $\widehat{\omega}$'s differ from 1 merely by chance? You may test the hypothesis of no association by using Fisher's exact test (for small samples) or the chi-squared test (for large samples).

For the three examples, using the table of χ^2 critical values with one degree of freedom, we test the statistical significance of the association by using the chi-square statistic with continuity correction.

Smoking–perinatal mortality:

$$X_c^2 = \frac{48{,}378[|619 \times 26{,}682 - 634 \times 20{,}443| - \frac{1}{2}(48{,}378)]^2}{21{,}062(27{,}316)(1253)(47{,}125)} = 17.76$$

From Table A.3, $p < 0.001$, and there is significant association. (Equivalently, for one degree of freedom, $Z = \sqrt{\chi_c^2} = 4.21$ and Table A.3 shows $p < 0.0001$.)

Breast cancer and screening:

$$X_c^2 = \frac{62{,}000[|40 \times 30{,}937 - 63 \times 30{,}960| - \frac{1}{2}(62{,}000)]^2}{31{,}000(31{,}000)(103)(61{,}897)} = 4.71$$

From the table, $0.01 < p < 0.05$ and the association is statistically significant at the 0.05 level.

Motor-vehicle job and herniated disk: $X_c^2 = 1.21$. From the χ^2 table, $p > 0.25$, and there is *not* a statistical association using only the Yale–New Haven data. In the next section we return to this data set.

If there is association, what can one say about the accuracy of the estimates? For the first two examples, where there is a statistically significant association, we turn to the construction of confidence intervals for ω. The procedure is to construct a confidence interval for $\ln \omega$, the natural log of ω, and to "exponentiate" the endpoints to find the confidence interval for ω. Our logarithms are natural logarithms, that is, to the base e. Recall e is a number; $e = 2.71828\ldots$.

The estimate of $\ln \omega$ is $\ln \widehat{\omega}$. The standard error of $\ln \widehat{\omega}$ is estimated by

$$\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

The estimate is approximately normally distributed; thus, normal critical values are used in constructing the confidence intervals. A $100(1-\alpha)\%$ confidence interval for $\ln \omega$ is given by

$$\ln \widehat{\omega} \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

where an $N(0, 1)$ variable has probability $\alpha/2$ of exceeding $z_{1-\alpha/2}$.

Upon finding the endpoints of this confidence interval, we exponentiate the values of the endpoints to find the confidence interval for ω. We find a 99% confidence interval for ω with the smoking and perinatal mortality data. First we construct the confidence interval for $\ln \omega$:

$$\ln(1.27) \pm 2.576\sqrt{\frac{1}{619} + \frac{1}{20{,}443} + \frac{1}{26{,}682} + \frac{1}{634}}$$

or 0.2390 ± 0.1475 or (0.0915, 0.3865). The confidence interval for ω is

$$(e^{0.0915}, e^{0.3865}) = (1.10, 1.47)$$

To find a 95% confidence interval for the breast cancer–screening data,

$$\ln(0.63) \pm 1.96\sqrt{\frac{1}{40} + \frac{1}{30{,}960} + \frac{1}{30{,}937} + \frac{1}{63}}$$

or -0.4620 ± 0.3966 or $(-0.8586, -0.0654)$. The 95% confidence interval for the odds ratio, ω, is (0.424, 0.937).

The reason for using logarithms in constructing the confidence intervals is that $\ln \widehat{\omega}$ is more normally distributed than ω. The standard error of ω may be estimated directly by

$$\widehat{\omega}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

(see Note 6.2 for the rationale). However, confidence intervals should be constructed as illustrated above.

6.3.5 Combination of 2 × 2 Tables

In this section we consider methods of combining 2×2 tables. The tables arise in one of two ways. In the first situation, we are interested in investigating an association between disease and exposure. There is, however, a third variable taking a finite number of values. We wish to "adjust" for the effect of the third variable. The values of the "confounding" third variable sometimes arise by taking a continuous variable and grouping by intervals; thus, the values are sometimes called *strata*. A second situation in which we will deal with several 2×2 tables is when the study of association and disease is made in more than one group. In some reasonable way, one would like to consider the combination of the 2×2 tables from each group.

Why Combine 2 × 2 *Tables?*

To see why one needs to worry about such things, suppose that there are two strata. In our first example there is no association between exposure and disease in each stratum, but if we ignore strata and "pool" our data (i.e., add it all together), an association appears. For stratum 1,

	Disease	
Exposure	+	−
+	5	50
−	10	100

$$\widehat{\omega}_1 = \frac{5(100)}{10(50)} = 1$$

and for stratum 2,

	Disease	
Exposure	**+**	**−**
+	40	60
−	40	60

$$\widehat{\omega}_2 = \frac{40(60)}{40(60)} = 1$$

In both tables the odds ratio is 1 and there is no association. Combining tables, the combined table and its odds ratio are:

	Disease	
Exposure	**+**	**−**
+	45	110
−	50	160

$$\widehat{\omega}_{\text{combined}} = \frac{45(160)}{50(110)} \doteq 1.31$$

When combining tables with no association, or odds ratios of 1, the combination may show association. For example, one would expect to find a positive relationship between breast cancer and being a homemaker. Possibly tables given separately for each gender would not show such an association. If the inference to be derived were that homemaking might be related causally to breast cancer, it is clear that one would need to adjust for gender.

On the other hand, there can be an association within each stratum that disappears in the pooled data set. The following numbers illustrate this:

Stratum 1:

	Disease	
Exposure	**+**	**−**
+	60	100
−	10	50

$$\widehat{\omega}_1 = \frac{60(50)}{10(100)} = 3$$

Stratum 2:

	Disease	
Exposure	**+**	**−**
+	50	10
−	100	60

$$\widehat{\omega}_2 = \frac{50(60)}{100(10)} = 3$$

Combined data:

	Disease	
Exposure	**+**	**−**
+	110	110
−	110	110

$$\widehat{\omega}_{\text{combined}} = 1$$

Thus, ignoring a confounding variable may "hide" an association that exists within each stratum but is not observed in the combined data.

Formally, our two situations are the same if we identify the stratum with differing groups. Also, note that there may be more than one confounding variable, that each strata of the "third" variable could correspond to a different combination of several other variables.

Questions of Interest in Multiple 2 × 2 *Tables*

In examining more than one 2×2 table, one or more of three questions is usually asked. This is illustrated by using the data of the study involving cases of acute herniated lumbar disk and controls (not matched) in Example 6.15, which compares the proportions with jobs driving motor vehicles. Seven different hospital services are involved, although only one of them was presented in Example 6.15. Numbering the sources from 1 to 7 and giving the data as 2×2 tables, the tables and the seven odds ratios are:

Source 1:

	Herniated Disk	
Motor Vehicle Job	**+**	**−**
+	8	1
−	47	26

$\widehat{\omega} = 4.43$

Source 2:

	+	−
+	5	0
−	17	21

$\widehat{\omega} = \infty$

Source 3:

	+	−
+	4	4
−	13	77

$\widehat{\omega} = 5.92$

Source 4:

	+	−
+	2	10
−	12	65

$\widehat{\omega} = 1.08$

Source 5:

	+	−
+	1	3
−	5	10

$\widehat{\omega} = 0.67$

Source 6:

	+	−
+	1	2
−	3	11

$\widehat{\omega} = 1.83$

Source 7:

	+	−
+	2	2
−	12	37

$\widehat{\omega} = 3.08$

The seven odds ratios are 4.43, ∞, 5.92, 1.08, 0.67, 1.83, and 3.08. The ratios vary so much that one might wonder whether each hospital service has the same degree of association (question 1). If they do not have the same degree of association, one might question whether the controls are appropriate, the patient populations are different, and so on.

One would also like an estimate of the overall or average association (question 2). From the previous examples it is seen that it might not be wise to sum all the tables and compute the association based on the pooled tables.

Finally, another question, related to the first two, is whether there is any evidence of any association, either overall or in some of the groups (question 3).

Two Approaches to Estimating an Overall Odds Ratio

If the seven different tables come from populations with the same odds ratio, how do we estimate the common or overall odds ratio? We will consider two approaches.

The first technique is to work with the natural logarithm, log to the base e, of the estimated odds ratio, $\widehat{\omega}$. Let $a_i = \ln \widehat{\omega}_i$, where $\widehat{\omega}_i$ is the estimated odds ratio in the ith of k 2×2 tables. The standard error of a_i is estimated by

$$s_i = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

where n_{11}, n_{12}, n_{21}, and n_{22} are the values from the ith 2×2 table. How do we investigate the problems mentioned above? To do this, one needs to understand a little of how the χ^2 distribution arises. The square of a standard normal variable has a chi-square distribution with one degree of freedom. If independent chi-square variables are added, the result is a chi-square variable whose degrees of freedom comprises the sum of the degrees of freedom of the variables that were added (see Note 5.3 also).

We now apply this to the problem at hand. Under the null hypothesis of no association in any of the tables, each a_i/s_i is approximately a standard normal value. If there is no association, $\omega = 1$ and $\ln \omega = 0$. Thus, $\log \widehat{\omega}_i$ has a mean of approximately zero. Its square, $(a_i/s_i)^2$, is approximately a χ^2 variable with one degree of freedom. The sum of all k of these independent, approximately chi-square variables is approximately a chi-square variable with k degrees of freedom. The sum is

$$X^2 = \sum_{i=1}^{k} \left(\frac{a_i}{s_i}\right)^2$$

and under the null hypothesis it has approximately a χ^2-distribution with k degrees of freedom.

It is possible to partition this sum into two parts. One part tests whether the association might be the same in all k tables (i.e., it tests for homogeneity). The second part will test to see whether on the basis of all the tables there is any association.

Suppose that one wants to "average" the association from all of the 2×2 tables. It seems reasonable to give more weight to the better estimates of association; that is, one wants the estimates with higher variances to get less weight. An appropriate weighted average is

$$\overline{a} = \sum_{i=1}^{k} \frac{a_i}{s_i^2} \Big/ \sum_{i=1}^{k} \frac{1}{s_i^2}$$

The χ^2-statistic then is partitioned, or broken down, into two parts:

$$X^2 = \sum_{i=1}^{k} \left(\frac{a_i}{s_i}\right)^2 = \sum_{i=1}^{k} \frac{1}{s_i^2}(a_i - \overline{a})^2 + \sum_{i=1}^{k} \frac{1}{s_i^2}\overline{a}^2$$

On the right-hand side, the first sum is approximately a χ^2 random variable with $k-1$ degrees of freedom if all k groups have the same degree of association. It tests for the homogeneity of the association in the different groups. That is, if χ^2 for homogeneity is too large, we reject the null hypothesis that the degree of association (whatever it is) is the same in each group. The second term tests whether there is association on the average. This has approximately a χ^2-distribution with one degree of freedom if there is no association in each group. Thus, define

$$\chi_H^2 = \sum_{i=1}^{k} \frac{1}{s_i^2}(a_i - \overline{a})^2 = \sum_{i=1}^{k} \frac{a_i^2}{s_i^2} - \overline{a}^2 \sum_{i=1}^{k} \frac{1}{s_i^2}$$

and

$$\chi_A^2 = \overline{a}^2 \sum_{i=1}^{k} \frac{1}{s_i^2}$$

Of course, if we decide that there are different degrees of association in different groups, this means that at least one of the groups must have some association.

Consider now the data given above. A few additional points are introduced. We use the log of the odds ratio, but the second group has $\widehat{\omega} = \infty$. What shall we do about this?

With small numbers, this may happen due to a zero in a cell. The bias of the method is reduced by adding 0.5 to each cell in each table:

[1]	+	−
+	8.5	1.5
−	47.5	26.5

[2]	+	−
+	5.5	0.5
−	17.5	21.5

[3]	+	−
+	4.5	4.5
−	13.5	77.5

[4]	+	−
+	2.5	10.5
−	12.5	65.5

[5]	+	−
+	1.5	3.5
−	5.5	10.5

[6]	+	−
+	1.5	2.5
−	3.5	11.5

[7]	+	−
+	2.5	2.5
−	12.5	37.5

Now

$$\widehat{\omega}_i = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)}, \qquad s_i = \sqrt{\frac{1}{n_{11} + 0.5} + \frac{1}{n_{22} + 0.5} + \frac{1}{n_{12} + 0.5} + \frac{1}{n_{21} + 0.5}}$$

The calculations above are shown in Table 6.3.

Table 6.3 Calculations for the Seven Tables

Table i	$\widehat{\omega}_i$	$a_i = \log \widehat{\omega}_i$	s_i^2	$1/s_i^2$	a_i^2/s_i^2	a_i/s_i^2
1	3.16	1.15	0.843	1.186	1.571	1.365
2	13.51	2.60	2.285	0.438	2.966	1.139
3	5.74	1.75	0.531	1.882	5.747	3.289
4	1.25	0.22	0.591	1.693	0.083	0.375
5	0.82	−0.20	1.229	0.813	0.033	−0.163
6	1.97	0.68	1.439	0.695	0.320	0.472
7	3.00	1.10	0.907	1.103	1.331	1.212
Total				7.810	12.051	7.689

Then

$$\overline{a} = \sum_{i=1}^{k} \frac{a_i}{s_i^2} \Big/ \sum_{i=1}^{k} \frac{1}{s_i^2} = \frac{7.689}{7.810} \doteq 0.985$$

$$X_A^2 = (0.985)^2(7.810) \doteq 7.57$$

$$X_H^2 = \sum \frac{a_i^2}{s_i^2} - \chi_A^2 = 12.05 - 7.57 = 4.48$$

X_H^2 with $7 - 1 = 6$ degrees of freedom has an $\alpha = 0.05$ critical value of 12.59 from Table A.3. We do *not* conclude that the association differs between groups.

Moving to the X_A^2, we find that $7.57 > 6.63$, the χ^2 critical value with one degree of freedom at the 0.010 level. We conclude that there *is* some overall association.

The odds ratio is estimated by $\widehat{\omega} = e^{\overline{a}} = e^{0.985} = 2.68$. The standard error of $\overline{a}$ is estimated by

$$\frac{1}{\sqrt{\sum_{i=1}^{k} (1/s_i^2)}}$$

To find a confidence interval for ω, first find one for $\ln \omega$ and "exponentiate" back. To find a 95% confidence interval, the calculation is

$$\overline{a} \pm \frac{z_{0.975}}{\sqrt{\sum (1/s_i^2)}} = 0.985 \pm \frac{1.96}{\sqrt{7.810}} \quad \text{or} \quad 0.985 \pm 0.701 \quad \text{or} \quad (0.284, 1.696)$$

Taking exponentials, the confidence interval for the overall odds ratio is (1.33, 5.45).

The second method of estimation is due to Mantel and Haenszel [1959]. Their estimate of the odds ratio is

$$\widehat{\omega} = \sum_{i=1}^{k} \frac{n_{11}(i)n_{22}(i)}{n_{..}(i)} \Big/ \sum_{i=1}^{k} \frac{n_{12}(i)n_{21}(i)}{n_{..}(i)}$$

where $n_{11}(i), n_{22}(i), n_{12}(i), n_{21}(i)$, and $n_{..}(i)$ are $n_{11}, n_{22}, n_{12}, n_{21}$, and $n_{..}$ for the ith table.

In this problem,

$$\widehat{\omega} = \frac{\dfrac{8 \times 26}{82} + \dfrac{5 \times 21}{43} + \dfrac{4 \times 77}{98} + \dfrac{2 \times 65}{89} + \dfrac{1 \times 10}{19} + \dfrac{1 \times 11}{17} + \dfrac{2 \times 37}{53}}{\dfrac{47 \times 1}{82} + \dfrac{17 \times 10}{43} + \dfrac{13 \times 4}{98} + \dfrac{12 \times 10}{89} + \dfrac{5 \times 3}{19} + \dfrac{3 \times 2}{17} + \dfrac{12 \times 12}{53}}$$

$$\doteq \frac{12.1516}{4.0473} \doteq 3.00$$

A test of association is given by the following statistic, X_A^2, which is approximately a chi-square random variable with one degree of freedom:

$$X_A^2 = \frac{\left[\left|\sum_{i=1}^{k} n_{11}(i) - \sum_{i=1}^{k} n_{1\cdot}(i)n_{\cdot 1}(i)/n_{..}(i)\right| - \frac{1}{2}\right]^2}{\sum_{i=1}^{k} n_{1\cdot}(i)n_{2\cdot}(i)n_{\cdot 1}(i)n_{\cdot 2}(i)/n_{..}(i)^2[n_{..}(i) - 1]}$$

The herniated disk data yield $X_A^2 = 7.92$, so that, as above, there is a significant ($p < 0.01$) association between an acute herniated lumbar intervertebral disk and whether or not a job

requires driving a motor vehicle. See Schlesselman [1982] and Breslow and Day [1980] for methods of setting confidence intervals for ω using the Mantel–Haenszel estimate.

In most circumstances, combining 2×2 tables will be used to adjust for other variables that define the strata (i.e., that define the different tables). The homogeneity of the odds ratio is usually of less interest unless the odds ratio differs widely among tables. Before testing for homogeneity of the odds ratio, one should be certain that this is what is desired (see Note 6.3).

6.3.6 Screening and Diagnosis: Sensitivity, Specificity, and Bayes' Theorem

In clinical medicine, and also in epidemiology, tests are often used to screen for the presence or absence of a disease. In the simplest case the test will simply be classified as having a positive (disease likely) or negative (disease unlikely) finding. Further, suppose that there is a "gold standard" that tells us whether or not a subject actually has the disease. The definitive classification might be based on data from follow-up, invasive radiographic or surgical procedures, or autopsy results. In many cases the gold standard itself will only be relatively correct, but nevertheless the best classification available. In this section we discuss summarization of the prediction of disease (as measured by our gold standard) by the test being considered. Ideally, those with the disease should all be classified as having disease, and those without disease should be classified as nondiseased. For this reason, two indices of the performance of a test consider how often such correct classification occurs.

Definition 6.3. The *sensitivity* of a test is the percentage of people with disease who are classified as having disease. A test is sensitive to the disease if it is positive for most people having the disease. The *specificity* of a test is the percentage of people without the disease who are classified as not having the disease. A test is specific if it is positive for a small percentage of those without the disease.

Further terminology associated with screening and diagnostic tests are true positive, true negative, false positive, and false negative tests.

Definition 6.4. A test is a *true positive test* if it is positive and the subject has the disease. A test is a *true negative test* if the test is negative and the subject does not have the disease. A *false positive test* is a positive test of a person without the disease. A *false negative test* is a negative test of a person with the disease.

Definition 6.5. The *predictive value of a positive test* is the percentage of subjects with a positive test who have the disease; the *predictive value of a negative test* is the percentage of subjects with a negative test who do not have the disease.

Suppose that data are collected on a test and presented in a 2×2 table as follows:

	Disease Category	
Screening Test Result	**Disease (+)**	**Nondiseased (−)**
Positive (+) test	a (true $+'$ s)	b (false $+'$ s)
Negative (−) test	c (false $-'$ s)	d (true $-'$ s)

The sensitivity is estimated by $100a/(a+c)$, the specificity by $100d/(b+d)$. If the subjects are representative of a population, the predictive value of positive and negative tests are estimated

by $100a/(a+b)$ and $100d/(c+d)$, respectively. These predictive values are useful only when the proportions with and without the disease in the study group are approximately the same as in the population where the test will be used to predict or classify (see below).

Example 6.16. Remein and Wilkerson [1961] considered a number of screening tests for diabetes. They had a group of consultants establish criteria, their gold standard, for diabetes. On each of a number of days, they recruited patients being seen in the outpatient department of the Boston City Hospital for reasons other than suspected diabetes. The table below presents results on the Folin–Wu blood test used 1 hour after a test meal and using a blood sugar level of 150 mg per 100 mL of blood sugar as a positive test.

Test	Diabetic	Nondiabetic	Total
+	56	49	105
−	14	461	475
Total	70	510	580

From this table note that there are 56 true positive tests compared to 14 false negative tests. The sensitivity is $100(56)/(56+14) = 80.0\%$. The 49 false positive tests and 461 true negative tests give a specificity of $100(461)/(49+461) = 90.4\%$. The predictive value of a positive test is $100(56)/(56+49) = 53.3\%$. The predictive value of a negative test is $100(461)/(14+461) = 97.1\%$.

If a test has a fixed value for its sensitivity and specificity, the predictive values will change depending on the prevalence of the disease in the population being tested. The values are related by *Bayes' theorem*. This theorem tells us how to update the probability of an event A: for example, the event of a subject having disease. If the subject is selected at random from some population, the probability of A is the fraction of people having the disease. Suppose that additional information becomes available; for example, the results of a diagnostic test might become available. In the light of this new information we would like to update or change our assessment of the probability that A occurs (that the subject has disease). The probability of A before receiving additional information is called the *a priori* or *prior probability*. The updated probability of A after receiving new information is called the *a posteriori* or *posterior probability*. Bayes' theorem is an explanation of how to find the posterior probability.

Bayes' theorem uses the concept of a conditional probability. We review this concept in Example 6.17.

Example 6.17. Comstock and Partridge [1972] conducted an informal census of Washington County, Maryland, in 1963. There were 127 arteriosclerotic heart disease deaths in the follow-up period. Of the deaths, 38 occurred among people whose usual frequency of church attendance was once or more per week. There were 24,245 such people as compared to 30,603 people whose usual attendance was less than once weekly. What is the probability of an arteriosclerotic heart disease death (event A) in three years given church attendance usually once or more per week (event B)?

From the data

$$P[A] = \frac{127}{24{,}245 + 30{,}603} = 0.0023$$

$$P[B] = \frac{24{,}245}{24{,}245 + 30{,}603} = 0.4420$$

$$P[A \ \& \ B] = \frac{38}{24{,}245 + 30{,}603} = 0.0007$$

$$P[A \mid B] = \frac{P[A \text{ and } B]}{P[B]} = \frac{0.0007}{0.4420} = 0.0016$$

If you knew that someone attended church once or more per week, the prior estimate of 0.0023 of the probability of an arteriosclerotic heart disease death in three years would be changed to a posterior estimate of 0.0016.

Using the conditional probability concept, Bayes' theorem may be stated.

Fact 1. (Bayes' Theorem) Let $B_1, \ldots, B_k$ be events such that one and only one of them must occur. Then for each i,

$$P[B_i|A] = \frac{P[A|B_i]P[B_i]}{P[A|B_1]P[B_1] + \cdots + P[A|B_k]P[B_k]}$$

Example 6.18. We use the data of Example 6.16 and Bayes' theorem to show that the predictive power of the test is related to the prevalence of the disease in the population. Suppose that the prevalence of the disease were not 70/580 (as in the data given), but rather, 6%. Also suppose that the sensitivity and specificity of the test were 80.0% and 90.4%, as in the example. What is the predictive value of a positive test?

We want $P[\text{disease}+|\text{test}+]$. Let B_1 be the event that the patient has disease and B_2 be the event of no disease. Let A be the occurrence of a positive test. A sensitivity of 80.0% is the same as $P[A|B_1] = 0.800$. A specificity of 90.4% is equivalent to $P[\text{not}A|B_2] = 0.904$. It is easy to see that

$$P[\text{not } A|B] + P[A|B] = 1$$

for any A and B. Thus, $P[A|B_2] = 1 - 0.904 = 0.096$. By assumption, $P[\text{disease}+] = P[B_1] = 0.06$, and $P[\text{disease}-] = P[B_2] = 0.94$. By Bayes' theorem,

$$P[\text{disease}+|\text{test}+] = \frac{P[\text{test} + |\text{disease}+]P[\text{disease}+]}{P[\text{test} + |\text{disease}+]P[\text{disease}+] + P[\text{test} + |\text{disease}-]P[\text{disease}-]}$$

Using our definitions of A, B_1, and B_2, this is

$$\begin{aligned} P[B_1|A] &= \frac{P[A|B_1]P[B_1]}{P[A|B_1]P[B_1] + P[A|B_2]P[B_2]} \\ &= \frac{0.800 \times 0.06}{0.800 \times 0.06 + 0.096 \times 0.94} \\ &= 0.347 \end{aligned}$$

If the disease prevalence is 6%, the predictive value of a positive test is 34.7% rather than 53.3% when the disease prevalence is 70/580 (12.1%).

Problems 6.15 and 6.28 illustrate the importance of disease prevalence in assessing the results of a test. See Note 6.8 for relationships among sensitivity, specificity, prevalence, and predictive values of a positive test. Sensitivity and specificity are discussed further in Chapter 13. See also Pepe [2003] for an excellent overview.

6.4 MATCHED OR PAIRED OBSERVATIONS

The comparisons among proportions in the preceding sections dealt with samples from different populations or from different subsets of a specified population. In many situations, the estimates of the proportions are based on the same objects or come from closely related, matched, or paired observations. You have seen matched or paired data used with a one-sample t-test.

A standard epidemiological tool is the retrospective paired case–control study. An example was given in Chapter 1. Let us recall the rationale for such studies. Suppose that one wants to see whether or not there is an association between a risk factor (say, use of oral contraceptives), and a disease (say, thromboembolism). Because the incidence of the disease is low, an extremely large prospective study would be needed to collect an adequate number of cases. One strategy is to *start* with the cases. The question then becomes one of finding appropriate controls for the cases. In a matched pair study, one control is identified for each case. The control, not having the disease, should be identical to the case in all relevant ways except, possibly, for the risk factor (see Note 6.6).

Example 6.19. This example is a retrospective matched pair case–control study by Sartwell et al. [1969] to study thromboembolism and oral contraceptive use. The cases were 175 women of reproductive age (15 to 44), discharged alive from 43 hospitals in five cities after initial attacks of idiopathic (i.e., of unknown cause) thrombophlebitis (blood clots in the veins with inflammation in the vessel walls), pulmonary embolism (a clot carried through the blood and obstructing lung blood flow), or cerebral thrombosis or embolism. The controls were matched with their cases for hospital, residence, time of hospitalization, race, age, marital status, parity, and pay status. More specifically, the controls were female patients from the same hospital during the same six-month interval. The controls were within five years of age and matched on parity (0, 1, 2, 3, or more prior pregnancies). The hospital pay status (ward, semiprivate, or private) was the same. The data for oral contraceptive use are:

	Control Use?	
Case Use?	**Yes**	**No**
Yes	10	57
No	13	95

The question of interest: Are cases more likely than controls to use oral contraceptives?

6.4.1 Matched Pair Data: McNemar's Test and Estimation of the Odds Ratio

The 2×2 table of Example 6.19 does not satisfy the assumptions of previous sections. The proportions using oral contraceptives among cases and controls cannot be considered samples from two populations since the cases and controls are paired; that is, they come together. Once a case is selected, the control for the case is constrained to be one of a small subset of people who match the case in various ways.

Suppose that there is no association between oral contraceptive use and thromboembolism after taking into account relevant factors. Suppose a case and control are such that only one of the pair uses oral contraceptives. Which one is more likely to use oral contraceptives? They may both be likely or unlikely to use oral contraceptives, depending on a variety of factors. Since the pair have the same values of such factors, neither member of the pair is more likely to have the risk factor! That is, in the case of disagreement, or discordant pairs, the probability that the case has the risk factor is 1/2. More generally, suppose that the data are

	Control Has Risk Factor?	
Case Has Risk Factor?	**Yes**	**No**
Yes	a	b
No	c	d

If there is no association between disease (i.e., case or control) and the presence or absence of the risk factor, the number b is binomial with $\pi = 1/2$ and $n = b + c$. To test for association we test $\pi = 1/2$, as shown previously. For large n, say $n \geq 30$,

$$X^2 = \frac{(b-c)^2}{b+c}$$

has a chi-square distribution with one degree of freedom if $\pi = 1/2$. For Example 6.19,

$$X^2 = \frac{(57-13)^2}{57+13} = 27.66$$

From the chi-square table, $p < 0.001$, so that there is a statistically significant association between thromboembolism and oral contraceptive use. This statistical test is called *McNemar's test*.

Procedure 6. For retrospective matched pair data, the odds ratio is estimated by

$$\widehat{\omega}_{\text{paired}} = \frac{b}{c}$$

The standard error of the estimate is estimated by

$$(1 + \widehat{\omega}_{\text{paired}})\sqrt{\frac{\widehat{\omega}_{\text{paired}}}{b+c}}$$

In Example 6.19, we estimate the odds ratio by

$$\widehat{\omega} = \frac{57}{13} \doteq 4.38$$

The standard error is estimated by

$$(1 + 4.38)\sqrt{\frac{4.38}{70}} \doteq 1.35$$

An approximate 95% confidence interval is given by

$$4.38 \pm (1.96)(1.35) \quad \text{or} \quad (1.74, 7.02)$$

More precise intervals may be based on the use of confidence intervals for a binomial proportion and the fact that $\widehat{\omega}_{\text{paired}}/(\widehat{\omega}_{\text{paired}} + 1) = b/(b+c)$ is a binomial proportion (see Fleiss [1981]). See Note 6.5 for further discussion of the chi-square analysis of paired data.

6.5 POISSON RANDOM VARIABLES

The Poisson distribution occurs primarily in two closely related situations. The first is a situation in which one counts discrete events in space or time, or some other continuous situation. For example, one might note the time of arrival (considered as a particular point in time) at an emergency medical service over a fixed time period. One may count the number of discrete occurrences of arrivals over this continuum of time. Conceptually, we may get any nonnegative integer, no matter how large, as our answer. A second example occurs when counting numbers of red blood cells that occur in a specified rectangular area marked off in the field of view. In a diluted blood sample where the distance between cells is such that they do not tend to "bump into each other," we may idealize the cells as being represented by points in the plane. Thus, within the particular area of interest, we are counting the number of points observed. A third example where one would expect to model the number of counts by a Poisson distribution would be a situation in which one is counting the number of particle emissions from a radioactive source. If the time period of observation is such that the radioactivity of the source does not decrease significantly (i.e., the time period is small compared to the half-life of a particle), the counts (which may be considered as coming at discrete time points) would again be modeled appropriately by a Poisson distribution.

The second major use of the Poisson distribution is as an approximation to the binomial distribution. If n is large and π is small in a binomial situation, the number of successes is very closely modeled by the Poisson distribution. The closeness of the approximation is specified by a mathematical theorem. As a rough rule of thumb, for most purposes the Poisson approximation will be adequate if π is less than or equal to 0.1 and n is greater than or equal to 20.

For the Poisson distribution to be an appropriate model for counting discrete points occurring in some sort of a continuum, the following two assumptions must hold:

1. The number of events occurring in one part of the continuum should be statistically independent of the number of events occurring in another part of the continuum. For example, in the emergency room, if we measure the number of arrivals during the first half hour, this event could reasonably be considered statistically independent of the number of arrivals during the second half hour. If there has been some cataclysmic event such as an earthquake, the assumption will not be valid. Similarly, in counting red blood cells in a diluted blood solution, the number of red cells in one square might reasonably be modeled as statistically independent of the number of red cells in another square.
2. The expected number of counts in a given part of the continuum should approach zero as its size approaches zero. Thus, in observing blood cells, one does not expect to find any in a very small area of a diluted specimen.

6.5.1 Examples of Poisson Data

Example 6.3 [Bucher et al., 1976] examines racial differences in the incidence of ABO hemolytic disease by examining records for infants born at the North Carolina Memorial Hospital. The samples of black and white infants gave the following estimated proportions with hemolytic disease:

$$\text{black infants}, \quad n_1 = 3584, \quad p_1 = 43/3584$$

$$\text{white infants}, \quad n_2 = 3831, \quad p_2 = 17/3831$$

The observed number of cases might reasonably be modeled by the Poisson distribution. (*Note:* The n is large and π is small in a binomial situation.) In this paper, studying the incidence of ABO hemolytic disease in black and white infants, the observed fractions for black and white infants of having the disease were 43/3584 and 17/3831. The 43 and 17 cases may be considered values of Poisson random variables.

A second example that would be modeled appropriately by the Poisson distribution is the number of deaths resulting from a large-scale vaccination program. In this case, n will be very large and π will be quite small. One might use the Poisson distribution in investigating the simultaneous occurrence of a disease and its association within a vaccination program. How likely is it that the particular "chance occurrence" might actually occur by chance?

Example 6.20. As a further example, a paper by Fisher et al. [1922] considers the accuracy of the plating method of estimating the density of bacterial populations. The process we are speaking about consists in making a suspension of a known mass of soil in a known volume of salt solution, and then diluting the suspension to a known degree. The bacterial numbers in the diluted suspension are estimated by plating a known volume in a nutrient gel medium and counting the number of colonies that develop from the plate. The estimate was made by a calculation that takes into account the mass of the soil taken and the degree of dilution. If we consider the colonies to be points occurring in the volume of gel, a Poisson model for the number of counts would be appropriate. Table 6.4 provides counts from seven different plates with portions of soil taken from a sample of Barnfield soil assayed in four parallel dilutions:

Example 6.21. A famous example of the Poisson distribution is data by von Bortkiewicz [1898] showing the chance of a cavalryman being killed by a horse kick in the course of a year (Table 6.5). The data are from recordings of 10 corps over a period of 20 years supplying 200 readings. A question of interest here might be whether a Poisson model is appropriate. Was the corps with four deaths an "unlucky" accident, or might there have been negligence of some kind?

Table 6.4 Counts for Seven Soil Samples

	Dilution			
Plate	I	II	III	IV
1	72	74	78	69
2	69	72	74	67
3	63	70	70	66
4	59	69	58	64
5	59	66	58	62
6	53	58	56	58
7	51	52	56	54
Mean	60.86	65.86	64.29	62.86

Table 6.5 Horse-kick Fatality Data

Number of Deaths per Corps per Year	Frequency
0	109
1	65
2	22
3	3
4	1
5	0
6	0

6.5.2 Poisson Model

The Poisson probability distribution is characterized by one parameter, λ. For each nonnegative integer k, if Y is a variable with the Poisson distribution with parameter λ,

$$P[Y = k] = \frac{e^{-\lambda}\lambda^k}{k!}$$

The parameter λ is both the mean and variance of the Poisson distribution,

$$E(Y) = \text{var}(Y) = \lambda$$

Bar graphs of the Poisson probabilities are given in Figure 6.3 for selected values of λ. As the mean (equal to the variance) increases, the distribution moves to the right and becomes more spread out and more symmetrical.

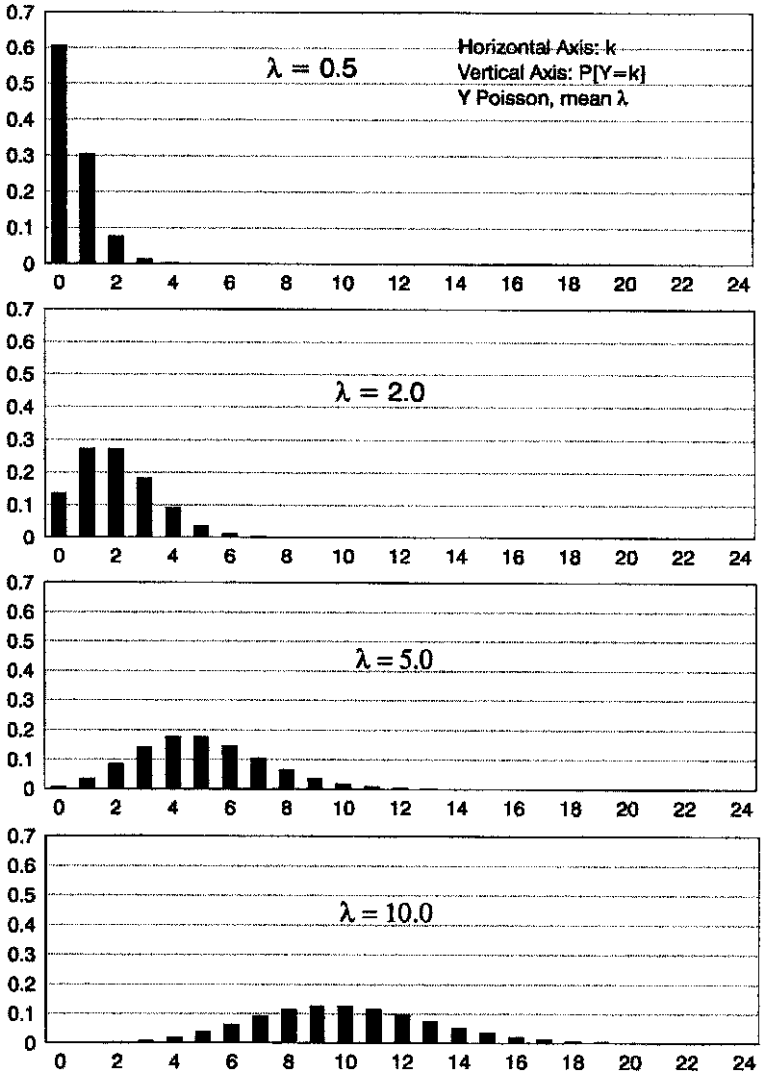


Figure 6.3 Poisson distribution.

Table 6.6 Binomial and Poisson Probabilities

	Binomial Probabilities			
k	$n = 10$ $\pi = 0.20$	$n = 20$ $\pi = 0.10$	$n = 40$ $\pi = 0.05$	Probabilities Poisson
0	0.1074	0.1216	0.1285	0.1353
1	0.2684	0.2702	0.2706	0.2707
2	0.3020	0.2852	0.2777	0.2707
3	0.2013	0.1901	0.1851	0.1804
4	0.0881	0.0898	0.0901	0.0902
5	0.0264	0.0319	0.0342	0.0361
6	0.0055	0.0089	0.0105	0.0120

In using the Poisson distribution to approximate the binomial distribution, the parameter λ is chosen to equal $n\pi$, the expected value of the binomial distribution. Poisson and binomial probabilities are given in Table 6.6 for comparison. This table gives an idea of the accuracy of the approximation (table entry is $P[Y = k], \lambda = 2 = n\pi$) for the first seven values of three distributions.

A fact that is often useful is that a sum of independent Poisson variables is itself a Poisson variable. The parameter for the sum is the sum of the individual parameter values. The parameter λ of the Poisson distribution is estimated by the sample mean when a sample is available. For example, the horse-kick data leads to an estimate of λ—say l—given by

$$l = \frac{0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1}{109 + 65 + 22 + 3 + 1} = 0.61$$

Now, we consider the construction of confidence intervals for a Poisson parameter. Consider the case of one observation, Y, and a small result, say, $Y \leq 100$. Note 6.8 describes how confidence intervals are calculated and there is a table in the Web appendix to this chapter. From this we find a 95% confidence interval for the proportion of black infants having ABO hemolytic disease, in the Bucher et al. [1976] study. The approximate Poisson variable is the binomial variable, which in this case is equal to 43; thus, a 95% confidence interval for $\lambda = n\pi$ is (31.12, 57.92). The equation $\lambda = n\pi$ equates the mean values for the Poisson and binomial models. Now $n\pi$ is in (31.12, 57.92) if and only if π is in the interval

$$\left(\frac{31.12}{n}, \frac{57.92}{n}\right)$$

In this case, $n = 3584$, so the confidence interval is

$$\left(\frac{31.12}{3584}, \frac{57.92}{3584}\right) \quad \text{or} \quad (0.0087, 0.0162)$$

These results are comparable with the 95% binomial limits obtained in Example 6.9: (0.0084, 0.0156).

6.5.3 Large-Sample Statistical Inference for the Poisson Distribution

Normal Approximation to the Poisson Distribution

The Poisson distribution has the property that the mean and variance are equal. For the mean large, say ≥ 100, the normal approximation can be used. That is, let $Y \sim$ Poisson(λ) and $\lambda \geq 100$. Then, approximately, $Y \sim N(\lambda, \lambda)$. An approximate $100(1 - \alpha)\%$ confidence interval

for λ can be formed from

$$Y \pm z_{1-\alpha/2}\sqrt{Y}$$

where $z_{1-\alpha/2}$ is a standard normal deviate at two-sided significance level α. This formula is based on the fact that Y estimates the mean as well as the variance. Consider, again, the data of Bucher et al. [1976] (Example 6.3) dealing with the incidence of ABO hemolytic disease. The observed value of Y, the number of black infants with ABO hemolytic disease, was 43. A 95% confidence interval for the mean, λ, is (31.12, 57.92). Even though $Y \leq 100$, let us use the normal approximation. The estimate of the variance, σ^2, of the normal distribution is $Y = 43$, so that the standard deviation is 6.56. An approximate 95% confidence interval is $43 \pm (1.96)(6.56)$, producing (30.1, 55.9), which is close to the values (31.12, 57.92) tabled.

Suppose that instead of one Poisson value, there is a random sample of size n, $Y_1, Y_2, \ldots, Y_n$ from a Poisson distribution with mean λ. How should one construct a confidence interval for λ based on these data? The sum $Y = Y_1 + Y_2 + \cdots + Y_n$ is Poisson with mean $n\lambda$. Construct a confidence interval for $n\lambda$ as above, say (L, U). Then, an appropriate confidence interval for λ is $(L/n, U/n)$. Consider Example 6.20, which deals with estimating the bacterial density of soil suspensions. The results for sample I were 72, 69, 63, 59, 59, 53, and 51. We want to set up a 95% confidence interval for the mean density using the seven observations. For this example, $n = 7$.

$$Y = Y_1 + Y_2 + \cdots + Y_7 = 72 + 69 + \cdots + 51 = 426$$

A 95% confidence interval for 7λ is $426 \pm 1.96\sqrt{426}$.

$$L = 385.55, \qquad \frac{L}{7} = 55.1$$

$$U = 466.45, \qquad \frac{U}{7} = 66.6$$

$$\overline{Y} = 60.9$$

The 95% confidence interval is (55.1, 66.6).

Square Root Transformation

It is often considered a disadvantage to have a distribution with a variance not "stable" but dependent on the mean in some way, as, for example, the Poisson distribution. The question is whether there is a transformation, $g(Y)$, of the variable such that the variance is no longer dependent on the mean. The answer is "yes." For the Poisson distribution, it is the square root transformation. It can be shown for "reasonably large" λ, say $\lambda \geq 30$, that if $Y \sim \text{Poisson}(\lambda)$, then $\text{var}(\sqrt{Y}) \doteq 0.25$.

A side benefit is that the distribution of $\sqrt{Y}$ is more "nearly normal," that is, for specified λ, the difference between the sampling distribution of $\sqrt{Y}$ and the normal distribution is smaller for most values than the difference between the distribution of Y and the normal distribution.

For the situation above, it is approximately true that

$$\sqrt{Y} \sim N(\sqrt{\lambda}, 0.25)$$

Consider Example 6.20 again. A confidence interval for $\sqrt{\lambda}$ will be constructed and then converted to an interval for λ. Let $X = \sqrt{Y}$.

Y	72	69	63	59	59	53	51
$X = \sqrt{Y}$	8.49	8.31	7.94	7.68	7.68	7.28	7.14

The sample mean and variance of X are $\overline{X} = 7.7886$ and $s_x^2 = 0.2483$. The sample variance is very close to the variance predicted by the theory $\sigma_x^2 = 0.2500$. A 95% confidence interval on $\sqrt{\lambda}$ can be set up from

$$\overline{X} \pm 1.96\frac{s_x}{\sqrt{7}} \quad \text{or} \quad 7.7886 \pm (1.96)\sqrt{\frac{0.2483}{7}}$$

producing lower and upper limits in the X scale.

$$L_x = 7.4195, \quad U_x = 8.1577$$
$$L_x^2 = 55.0, \quad U_x^2 = 66.5$$

which are remarkably close to the values given previously.

Poisson Homogeneity Test

In Chapter 4 the question of a test of normality was discussed and a graphical procedure was suggested. Fisher et al. [1922], in the paper described in Example 6.20, derived an approximate test for determining whether or not a sample of observations could have come from a Poisson distribution with the same mean. The test does not determine "Poissonness," but rather, equality of means. If the experimental situations are identical (i.e., we have a random sample), the test is a test for Poissonness.

The test, the *Poisson homogeneity test*, is based on the property that for the Poisson distribution, the mean equals the variance. The test is the following: Suppose that $Y_1, Y_2, \ldots, Y_n$ are a random sample from a Poisson distribution with mean λ. Then, for a large λ—say, $\lambda \geq 50$—the quantity

$$X^2 = \frac{(n-1)s^2}{\overline{Y}}$$

has approximately a chi-square distribution with $n - 1$ degrees of freedom, where s^2 is the sample variance.

Consider again the data in Example 6.20. The mean and standard deviation of the seven observations are

$$n = 7, \qquad \overline{Y} = 60.86, \qquad s_y = 7.7552$$
$$X^2 = \frac{(7-1)(7.7552)^2}{60.86} = 5.93$$

Under the null hypothesis that all the observations are from a Poisson distribution with the same mean, the statistic $X^2 = 5.93$ can be referred to a chi-square distribution with six degrees of freedom. What will the rejection region be? This is determined by the alternative hypothesis. In this case it is reasonable to suppose that the sample variance will be greater than expected if the null hypothesis is not true. Hence, we want to reject the null hypothesis when χ^2 is "large"; "large" in this case means $P[X^2 \geq \chi_{1-\alpha}^2] = \alpha$.

Suppose that $\alpha = 0.05$; the critical value for $\chi_{1-\alpha}^2$ with 6 degrees of freedom is 12.59. The observed value $X^2 = 5.93$ is much less than that and the null hypothesis is not rejected.

6.6 GOODNESS-OF-FIT TESTS

The use of appropriate mathematical models has made possible advances in biomedical science; the key word is *appropriate*. An inappropriate model can lead to false or inappropriate ideas.

In some situations the appropriateness of a model is clear. A random sample of a population will lead to a binomial variable for the response to a yes or no question. In other situations the issue may be in doubt. In such cases one would like to examine the data to see if the model used seems to fit the data. Tests of this type are called *goodness-of-fit tests*. In this section we examine some tests where the tests are based on count data. The count data may arise from continuous data. One may count the number of observations in different intervals of the real line; examples are given in Sections 6.6.2 and 6.6.4.

6.6.1 Multinomial Random Variables

Binomial random variables count the number of successes in n independent trials where one and only one of two possibilities must occur. *Multinomial random variables* generalize this to allow more than two possible outcomes. In a multinomial situation, outcomes are observed that take one and only one of two or more, say k, possibilities. There are n independent trials, each with the same probability of a particular outcome. Multinomial random variables count the number of occurrences of a particular outcome. Let n_i be the number of occurrences of outcome i. Thus, n_i is an integer taking a value among $0, 1, 2, \ldots, n$. There are k different n_i, which add up to n since one and only one outcome occurs on each trial:

$$n_1 + n_2 + \cdots + n_k = n$$

Let us focus on a particular outcome, say the ith. What are the mean and variance of n_i? We may classify each outcome into one of two possibilities, the ith outcome or anything else. There are then n independent trials with two outcomes. We see that n_i is a binomial random variable when considered alone. Let π_i, where $i = 1, \ldots, k$, be the probability that the ith outcome occurs. Then

$$E(n_i) = n\pi_i, \qquad \text{var}(n_i) = n\pi_i(1 - \pi_i) \tag{6}$$

for $i = 1, 2, \ldots, k$.

Often, multinomial outcomes are visualized as placing the outcome of each of the n trials into a separate *cell* or box. The probability π_i is then the probability that an outcome lands in the ith cell.

The remainder of this section deals with multinomial observations. Tests are presented to see if a specified multinomial model holds.

6.6.2 Known Cell Probabilities

In this section, the cell probabilities $\pi_1, \ldots, \pi_k$ are specified. We use the specified values as a null hypothesis to be compared with the data $n_1, \ldots, n_k$. Since $E(n_i) = n\pi_i$, it is reasonable to examine the differences $n_i - n\pi_i$. The statistical test is given by the following fact.

Fact 2. Let n_i, where $i = 1, \ldots, k$, be multinomial. Under $H_0 : \pi_i = \pi_i^0$,

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i^0)^2}{n\pi_i^0}$$

has approximately a chi-square distribution with $k - 1$ degrees of freedom. If some π_i are not equal to π_i^0, X^2 will tend to be too large.

The distribution of X^2 is well approximated by the chi-square distribution if all of the expected values, $n\pi_i^0$, are at least five, except possibly for one or two of the values. When the null hypothesis is not true, the null hypothesis is rejected for X^2 too large. At significance level

α, reject H_0 if $X^2 \geq \chi^2_{1-\alpha,k-1}$, where $\chi^2_{1-\alpha,k-1}$ is the $1-\alpha$ percentage point for a χ^2 random variable with $k-1$ degrees of freedom.

Since there are k cells, one might expect the labeling of the degrees of freedom to be k instead of $k-1$. However, since the n_i add up to n we only need to know $k-1$ of them to know all k values. There are really only $k-1$ quantities that may vary at a time; the last quantity is specified by the other $k-1$ values.

The form of X^2 may be kept in mind by noting that we are comparing the observed values, n_i, and expected values, $n\pi_i^0$. Thus,

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Example 6.22. Are births spread uniformly throughout the year? The data in Table 6.7 give the number of births in King County, Washington, from 1968 through 1979 by month. The estimated probability of a birth in a given month is found by taking the number of days in that month and dividing by the total number of days (leap years are included in Table 6.7).

Testing the null hypothesis using Table A.3, we see that $163.15 > 31.26 = \chi^2_{0.001,11}$, so that $p < 0.001$. We reject the null hypothesis that births occur uniformly throughout the year. With this large sample size ($n = 160{,}654$) it is not surprising that the null hypothesis can be rejected. We can examine the magnitude of the effect by comparing the ratio of observed to expected numbers of births, with the results shown in Table 6.8. There is an excess of births in the spring (March and April) and a deficit in the late fall and winter (October through January). Note that the difference from expected values is small. The maximum "excess" of births occurred

Table 6.7 Births in King County, Washington, 1968–1979

Month	Births	Days	π_i^0	$n\pi_i^0$	$(n_i - n\pi_i^0)^2/n\pi_i^0$
January	13,016	310	0.08486	13,633	27.92
February	12,398	283	0.07747	12,446	0.19
March	14,341	310	0.08486	13,633	36.77
April	13,744	300	0.08212	13,193	23.01
May	13,894	310	0.08486	13,633	5.00
June	13,433	300	0.08212	13,193	4.37
July	13,787	310	0.08486	13,633	1.74
August	13,537	310	0.08486	13,633	0.68
September	13,459	300	0.08212	13,193	5.36
October	13,144	310	0.08486	13,633	17.54
November	12,497	300	0.08212	13,193	36.72
December	13,404	310	0.08486	13,633	3.85
Total	160,654 (n)	3653	0.99997		$163.15 = X^2$

Table 6.8 Ratios of Observed to Expected Births

Month	Observed/Expected Births	Month	Observed/Expected Births
January	0.955	July	1.011
February	0.996	August	0.993
March	1.052	September	1.020
April	1.042	October	0.964
May	1.019	November	0.947
June	1.018	December	0.983

in March and was only 5.2% above the number expected. A plot of the ratio vs. month would show a distinct sinusoidal pattern.

Example 6.23. Mendel [1866] is justly famous for his theory and experiments on the principles of heredity. Sir R. A. Fisher [1936] reviewed Mendel's work and found a surprisingly good fit to the data. Consider two parents heterozygous for a dominant–recessive trait. That is, each parent has one dominant gene and one recessive gene. Mendel hypothesized that all four combinations of genes would be equally likely in the offspring. Let A denote the dominant gene and a denote the recessive gene. The two parents are Aa. The offspring should be

Genotype	Probability
AA	1/4
Aa	1/2
aa	1/4

The Aa combination has probability 1/2 since one cannot distinguish between the two cases where the dominant gene comes from one parent and the recessive gene from the other parent. In one of Mendel's experiments he examined whether a seed was wrinkled, denoted by a, or smooth, denoted by A. By looking at offspring of these seeds, Mendel classified the seeds as aa, Aa, or AA. The results were

	AA	Aa	aa	Total
Number	159	321	159	639

as presented in Table II of Fisher [1936]. Do these data support the hypothesized $1:2:1$ ratio? The chi-square statistic is

$$X^2 = \frac{(159-159.75)^2}{159.75} + \frac{(321-319.5)^2}{319.5} + \frac{(159-159.75)^2}{159.75} = 0.014$$

For the χ^2 distribution with two degrees of freedom, $p > 0.95$ from Table A.3 (in fact $p = 0.993$), so that the result has more agreement than would be expected by chance. We return to these data in Example 6.24.

6.6.3 Addition of Independent Chi-Square Variables: Mean and Variance of the Chi-Square Distribution

Chi-square random variables occur so often in statistical analysis that it will be useful to know more facts about chi-square variables. In this section facts are presented and then applied to an example (see also Note 5.3).

Fact 3. Chi-square variables have the following properties:

1. Let X^2 be a chi-square random variable with m degrees of freedom. Then

$$E(X^2) = m \quad \text{and} \quad \text{var}(X^2) = 2m$$

2. Let $X_1^2, \ldots, X_n^2$ be independent chi-square variables with $m_1, \ldots, m_n$ degrees of freedom. Then $X^2 = X_1^2 + \cdots + X_n^2$ is a chi-square random variable with $m = m_1 + m_2 + \cdots + m_n$ degrees of freedom.

Table 6.9 Chi-Square Values for Mendel's Experiments

Experiments	X^2	Degrees of Freedom
3 : 1 Ratios	2.14	7
2 : 1 Ratios	5.17	8
Bifactorial experiments	2.81	8
Gametic ratios	3.67	15
Trifactorial experiments	15.32	26
Total	29.11	64

3. Let X^2 be a chi-square random variable with m degrees of freedom. If m is large, say $m \geq 30$,

$$\frac{X^2 - m}{\sqrt{2m}}$$

is approximately a $N(0, 1)$ random variable.

Example 6.24. We considered Mendel's data, reported by Fisher [1936], in Example 6.23. As Fisher examined the data, he became convinced that the data fit the hypothesis too well [Box, 1978, pp. 195, 300]. Fisher comments: "Although no explanation can be expected to be satisfactory, it remains a possibility among others that Mendel was deceived by some assistant who knew too well what was expected."

One reason Fisher arrived at his conclusion was by combining χ^2 values from different experiments by Mendel. Table 6.9 presents the data.

If all the null hypotheses are true, by the facts above, $X^2 = 29.11$ should look like a χ^2 with 64 degrees of freedom. An approximate normal variable,

$$Z = \frac{29.11 - 64}{\sqrt{128}} = -3.08$$

has less than 1 chance in 1000 of being this small ($p = 0.99995$). One can only conclude that something peculiar occurred in the collection and reporting of Mendel's data.

6.6.4 Chi-Square Tests for Unknown Cell Probabilities

Above, we considered tests of the goodness of fit of multinomial data when the probability of being in an individual cell was specified precisely: for example, by a genetic model of how traits are inherited. In other situations, the cell probabilities are not known but may be estimated. First, we motivate the techniques by presenting a possible use; next, we present the techniques, and finally, we illustrate the use of the techniques by example.

Consider a sample of n numbers that may come from a normal distribution. How might we check the assumption of normality? One approach is to divide the real number line into a finite number of intervals. The number of points observed in each interval may then be counted. The numbers in the various intervals or cells are multinomial random variables. If the sample were normal with known mean μ and known standard deviation σ, the probability, π_i, that a point falls between the endpoints of the ith interval—say Y_1 and Y_2—is known to be

$$\pi_i = \Phi\left(\frac{Y_2 - \mu}{\sigma}\right) - \Phi\left(\frac{Y_1 - \mu}{\sigma}\right)$$

where Φ is the distribution function of a standard normal random variable. In most cases, μ and σ are not known, so μ and σ, and thus π_i, must be estimated. Now π_i depends on two variables, μ and σ : $\pi_i = \pi_i(\mu, \sigma)$ where the notation $\pi_i(\mu, \sigma)$ means that π_i is a function of μ and σ. It is natural if we estimate μ and σ by, say, $\widehat{\mu}$ and $\widehat{\sigma}$, to estimate π_i by $p_i(\widehat{\mu}, \widehat{\omega})$. That is,

$$p_i(\widehat{\mu}, \widehat{\sigma}) = \Phi\left(\frac{Y_2 - \widehat{\mu}}{\widehat{\sigma}}\right) - \Phi\left(\frac{Y_1 - \widehat{\mu}}{\widehat{\sigma}}\right)$$

From this, a statistic (X^2) can be formed as above. If there are k cells,

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{k} \frac{[n_i - np_i(\widehat{\mu}, \widehat{\sigma})]^2}{np_i(\widehat{\mu}, \widehat{\sigma})}$$

Does X^2 now have a chi-square distribution? The following facts describe the situation.

Fact 4. Suppose that n observations are grouped or placed into k categories or cells such that the probability of being in cell i is $\pi_i = \pi_i(\Theta_1, \ldots, \Theta_s)$, where π_i depends on s parameters Θ_j and where $s < k-1$. Suppose that none of the s parameters are determined by the remaining $s - 1$ parameters. Then:

1. If $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_s$, the parameter estimates, are chosen to minimize X^2, the distribution of X^2 is approximately a chi-square random variable with $k - s - 1$ degrees of freedom for large n. Estimates chosen to minimize the value of X^2 are called *minimum chi-square estimates*.
2. If estimates of $\Theta_1, \ldots, \Theta_s$ other than the minimum chi-square estimates are used, then for large n the distribution function of X^2 lies between the distribution functions of chi-square variables with $k - s - 1$ degrees of freedom and $k - 1$ degrees of freedom. More specifically, let $X^2_{1-\alpha,m}$ denote the α-significance-level critical value for a chi-square distribution with m degrees of freedom. The significance-level-α critical value of X^2 is less than or equal to $X^2_{1-\alpha,k-1}$. A conservative test of the multinomial model is to reject the null hypothesis that the model is correct if $X^2 \geq \chi^2_{1-\alpha,k-1}$.

These complex statements are best understood by applying them to an example.

Example 6.25. Table 3.4 in Section 3.3.1 gives the age in days at death of 78 SIDS cases. Test for normality at the 5% significance level using a χ^2-test.

Before performing the test, we need to divide the real number line into intervals or cells. The usual approach is to:

1. Estimate the parameters involved. In this case the unknown parameters are μ and σ. We estimate by $\overline{Y}$ and s.
2. Decide on k, the number of intervals. Let there be n observations. A good approach is to choose k as follows:
   - **a.** For $20 \leq n \leq 100, k \doteq n/5$.
   - **b.** For $n > 300, k \doteq 3.5n^{2/5}$ (here, $n^{2/5}$ is n raised to the 2/5 power).

3. Find the endpoints of the k intervals so that each interval has probability $1/k$. The k intervals are

$$\begin{array}{cc} (-\infty, a_1] & \text{interval 1} \\ (a_1, a_2] & \text{interval 2} \\ \vdots & \vdots \\ (a_{k-2}, a_{k-1}] & \text{interval } (k-1) \\ (a_{k-1}, \infty) & \text{interval k} \end{array}$$

Let Z_i be a value such that a standard normal random variable takes a value less than Z_i with probability i/k. Then

$$a_i = \overline{X} + sZ_i$$

(In testing for a distribution other than the normal distribution, other methods of finding cells of approximately equal probability need to be used.)

4. Compute the statistic

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n/k)^2}{n/k}$$

where n_i is the number of data points in cell i.

To apply steps 1 to 4 to the data at hand, one computes $n = 78$, $\overline{X} = 97.85$, and $s = 55.66$. As $78/5 = 15.6$, we will use $k = 15$ intervals. From tables of the normal distribution, we find Z_i, $i = 1, 2, \ldots, 14$, so that a standard normal random variable has probability $i/15$ of being less than Z_i. The values of Z_i and a_i are given in Table 6.10.

The number of observations observed in the 15 cells, from left to right, are 0, 8, 7, 5, 7, 9, 7, 5, 6, 6, 2, 2, 3, 5, and 6. In each cell, the number of observations expected is $np_i = n/k$ or $78/15 = 5.2$. Then

$$X^2 = \frac{(0-5.2)^2}{5.2} + \frac{(8-5.2)^2}{5.2} + \cdots + \frac{(6-5.2)^2}{5.2} = 16.62$$

We know that the 0.05 critical values are between the chi-square critical values with 12 and 14 degrees of freedom. The two values are 21.03 and 23.68. Thus, we do not reject the hypothesis of normality. (If the X^2 value had been greater than 23.68, we would have rejected the null hypothesis of normality. If X^2 were between 21.03 and 23.68, the answer would be in doubt. In that case, it would be advisable to compute the minimum chi-square estimates so that a known distribution results.)

Note that the largest observation, 307, is $(307 - 97.85)/55.6 = 3.76$ sample standard deviations from the sample mean. In using a chi-square goodness-of-fit test, all large observations are placed into a single cell. The magnitude of the value is lost. If one is worried about large outlying values, there are better tests of the fit to normality.

Table 6.10 Z_i and a_i Values

i	Z_i	a_i	i	Z_i	a_i	i	Z_i	a_i
1	−1.50	12.8	6	−0.25	84.9	11	0.62	135.0
2	−1.11	35.3	7	−0.08	94.7	12	0.84	147.7
3	−0.84	50.9	8	0.08	103.9	13	1.11	163.3
4	−0.62	63.5	9	0.25	113.7	14	1.50	185.8
5	−0.43	74.5	10	0.43	124.1			

NOTES

6.1 *Continuity Correction for* 2 × 2 *Table Chi-Square Values*

There has been controversy about the appropriateness of the continuity correction for 2×2 tables [Conover, 1974]. The continuity correction makes the *actual* significance levels under the null hypothesis closer to the hypergeometric (Fisher's exact test) actual significance levels. When compared to the chi-square distribution, the *actual* significance levels are too low [Conover, 1974; Starmer et al., 1974; Grizzle, 1967]. The *uncorrected* "chi-square" value referred to chi-square critical values gives actual and nominal significance levels that are close. For this reason, the authors recommend that the continuity correction *not* be used. Use of the continuity correction would be correct but overconservative. For arguments on the opposite side, see Mantel and Greenhouse [1968]. A good summary can be found in Little [1989].

6.2 *Standard Error of* $\widehat{\omega}$ *as Related to the Standard Error of* $\log \widehat{\omega}$

Let X be a positive variate with mean μ_x and standard deviation σ_x. Let $Y = \log_e X$. Let the mean and standard deviation of Y be μ_y and σ_y, respectively. It can be shown that under certain conditions

$$\frac{\sigma_x}{\mu_x} \doteq \sigma_y$$

The quantity σ_x/μ_x is known as the *coefficient of variation*. Another way of writing this is

$$\sigma_x \doteq \mu_x \sigma_y$$

If the parameters are replaced by the appropriate statistics, the expression becomes

$$s_x \doteq \overline{x} s_y$$

and the standard deviation of $\widehat{\omega}$ then follows from this relationship.

6.3 *Some Limitations of the Odds Ratio*

The odds ratio uses one number to summarize four numbers, and some information about the relationship is necessarily lost. The following example shows one of the limitations. Fleiss [1981] discusses the limitations of the odds ratio as a measure for public health. He presents the mortality rates per 100,000 person-years from lung cancer and coronary artery disease for smokers and nonsmokers of cigarettes [U.S. Department of Health, Education and Welfare, 1964]:

	Smokers	Nonsmokers	Odds Ratio	Difference
Cancer of the lung	48.33	4.49	10.8	43.84
Coronary artery disease	294.67	169.54	1.7	125.13

The point is that although the risk ω is increased much more for cancer, the added number dying of coronary artery disease is higher, and in some sense smoking has a greater effect in this case.

6.4 *Mantel–Haenszel Test for Association*

The chi-square test of association given in conjunction with the Mantel–Haenszel test discussed in Section 6.3.5 arises from the approach of the section by choosing a_i and s_i appropriately

[Fleiss, 1981]. The corresponding chi-square test for homogeneity does *not* make sense and should not be used. Mantel et al. [1977] give the problems associated with using this approach to look at homogeneity.

6.5 Matched Pair Studies

One of the difficult aspects in the design and execution of matched pair studies is to decide on the matching variables, and then to find matches to the degree desired. In practice, many decisions are made for logistic and monetary reasons; these factors are not discussed here. The primary purpose of matching is to have a *valid* comparison. Variables are matched to increase the validity of the comparison. Inappropriate matching can hurt the statistical power of the comparison. Breslow and Day [1980] and Miettinen [1970] give some fundamental background. Fisher and Patil [1974] further elucidate the matter (see also Problem 6.30).

6.6 More on the Chi-Square Goodness-of-Fit Test

The goodness-of-fit test as presented in this chapter did not mention some of the subtleties associated with the subject. A few arcane points, with appropriate references, are given in this note.

1. In Fact 4, the estimate used should be maximum likelihood estimates or equivalent estimates [Chernoff and Lehmann, 1954].
2. The initial chi-square limit theorems were proved for fixed cell boundaries. Limiting theorems where the boundaries were random (depending on the data) were proved later [Kendall and Stuart, 1967, Secs. 30.20 and 30.21].
3. The number of cells to be used (as a function of the sample size) has its own literature. More detail is given in Kendall and Stuart [1967, Secs. 30.28 to 30.30]. The recommendations for k in the present book are based on this material.

6.7 Predictive Value of a Positive Test

The predictive value of a positive test, PV^+, is related to the prevalence (PREV), sensitivity (SENS), and specificity (SPEC) of a test by the following equation:

$$\text{PV}^+ = \frac{1}{1 + \left[(1 - \text{SPEC})/\text{SENS}\right]\left[(1 - \text{PREV})/\text{PREV}\right]}$$

Here PREV, SENS, and SPEC, are on a scale of 0 to 1 of proportions instead of percentages.

If we define $\text{logit}(p) = \log[p/(1-p)]$, the predictive value of a positive test is related very simply to the prevalence as follows:

$$\text{logit}[\text{PV}^+] = \log\left(\frac{\text{SENS}}{1 - \text{SPEC}}\right) + \text{logit}(\text{PREV})$$

This is a very informative formula. For rare diseases (i.e., low prevalence), the term "logit (PREV)" will dominate the predictive value of a positive test. So no matter what the sensitivity or specificity of a test, the predictive value will be low.

6.8 Confidence Intervals for a Poisson Mean

Many software packages now provide confidence intervals for the mean of a Poisson distribution. There are two formulas: an approximate one that can be done by hand, and a more complex exact formula. The approximate formula uses the following steps. Given a Poisson variable Y:

1. Take $\sqrt{Y}$.
2. Add and subtract 1.
3. Square the result $[(\sqrt{Y}-1)^2,\ (\sqrt{Y}+1)^2]$.

This formula is reasonably accurate for $Y \geq 5$. See also Note 6.9 for a simple confidence interval when $Y = 0$. The exact formula uses the relationship between the Poisson and χ^2 distributions to give the confidence interval

$$\left[\frac{1}{2}\chi^2_{\alpha/2}(2x),\ \frac{1}{2}\chi^2_{1-\alpha/2}(2x+2)\right]$$

where $\chi^2_{\alpha/2}(2x)$ is the $\alpha/2$ percentile of the χ^2 distribution with $2x$ degrees of freedom.

6.9 Rule of Threes

An upper 90% confidence bound for a Poisson random variable with observed values 0 is, to a very good approximation, 3. This has led to the *rule of threes*, which states that if in n trials zero events of interest are observed, a 95% confidence bound on the underlying rate is $3/n$. For a fuller discussion, see Hanley and Lippman-Hard [1983]. See also Problem 6.29.

PROBLEMS

6.1 In a randomized trial of surgical and medical treatment a clinic finds eight of nine patients randomized to medicine. They complain that the randomization must not be working; that is, π cannot be 1/2.

(a) Is their argument reasonable from their point of view?

***(b)** With 15 clinics in the trial, what is the probability that *all* 15 clinics have fewer than eight people randomized to each treatment, of the first nine people randomized? Assume independent binomial distributions with $\pi = 1/2$ at each site.

6.2 In a dietary study, 14 of 20 subjects lost weight. If weight is assumed to fluctuate by chance, with probability 1/2 of losing weight, what is the exact two-sided p-value for testing the null hypothesis $\pi = 1/2$?

6.3 Edwards and Fraccaro [1960] present Swedish data about the gender of a child and the parity. These data are:

	Order of Birth							
Gender	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Males	2846	2554	2162	1667	1341	987	666	12,223
Females	2631	2361	1996	1676	1230	914	668	11,476
Total	5477	4915	4158	3343	2571	1901	1334	23,699

(a) Find the p-value for testing the hypothesis that a birth is equally likely to be of either gender using the combined data and binomial assumptions.

(b) Construct a 90% confidence interval for the probability that a birth is a female child.

(c) Repeat parts (a) and (b) using only the data for birth order 6.

6.4 Ounsted [1953] presents data about cases with convulsive disorders. Among the cases there were 82 females and 118 males. At the 5% significance level, test the hypothesis that a case is equally likely to be of either gender. The siblings of the cases were 121 females and 156 males. Test at the 10% significance level the hypothesis that the siblings represent 53% or more male births.

6.5 Smith et al. [1976] report data on ovarian carcinoma (cancer of the ovaries). People had different numbers of courses of chemotherapy. The five-year survival data for those with 1–4 and 10 or more courses of chemotherapy are:

	Five-Year Status	
Courses	**Dead**	**Alive**
1–4	21	2
$\geq$10	2	8

Using Fisher's exact test, is there a statistically significant association ($p \leq 0.05$) in this table? (In this problem and the next, you will need to compute the hypergeometric probabilities using the results of Problem 6.26.)

6.6 Borer et al. [1980] study 45 patients following an acute myocardial infarction (heart attack). They measure the *ejection fraction* (EF), the percent of the blood pumped from the left ventricle (the pumping chamber of the heart) during a heart beat. A low EF indicates damaged or dead heart muscle (myocardium). During follow-up, four patients died. Dividing EF into low (<35%) and high ($\geq$35%) EF groups gave the following table:

	Vital Status	
EF	**Dead**	**Alive**
<35%	4	9
$\geq$35%	0	32

Is there reason to suspect, at a 0.05 significance level, that death is more likely in the low EF group? Use a one-sided p-value for your answer, since biological plausibility (and prior literature) indicates that low EF is a risk factor for mortality.

6.7 Using the data of Problem 6.4, test the hypothesis that the proportions of male births among those with convulsive disorders and among their siblings are the same.

6.8 Lawson and Jick [1976] compare drug prescription in the United States and Scotland.

(a) In patients with congestive heart failure, two or more drugs were prescribed in 257 of 437 U.S. patients. In Scotland, 39 of 179 patients had two or more drugs prescribed. Test the null hypothesis of equal proportions giving the resulting p-value. Construct a 95% confidence interval for the difference in proportions.

(b) Patients with dehydration received two or more drugs in 55 of 74 Scottish cases as compared to 255 of 536 in the United States. Answer the questions of part (a).

6.9 A randomized study among patients with angina (heart chest pain) is to be conducted with five-year follow-up. Patients are to be randomized to medical and surgical treatment. Suppose that the estimated five-year medical mortality is 10% and it is hoped that the surgical mortality will be only half as much (5%) or less. If a test of binomial proportions at the 5% significance level is to be performed, and we want to be 90% certain of detecting a difference of 5% or more, what sample sizes are needed for the two (equal-sized) groups?

6.10 A cancer with poor prognosis, a three-year mortality of 85%, is studied. A new mode of chemotherapy is to be evaluated. Suppose that when testing at the 0.10 significance level, one wishes to be 95% certain of detecting a difference if survival has been increased to 50% or more. The randomized clinical trial will have equal numbers of people in each group. How many patients should be randomized?

6.11 Comstock and Partridge [1972] show data giving an association between church attendance and health. From the data of Example 6.17, which were collected from a prospective study:

(a) Compute the relative risk of an arteriosclerotic death in the three-year follow-up period if one usually attends church less than once a week as compared to once a week or more.

(b) Compute the odds ratio and a 95% confidence interval.

(c) Find the percent error of the odds ratio as an approximation to the relative risk; that is, compute $100(\text{OR} - \text{RR})/\text{RR}$.

(d) The data in this population on deaths from cirrhosis of the liver are:

Usual Church Attendance	Cirrhosis Fatality?	
	Yes	No
$\geq$1 per week	5	24,240
$<$1 per week	25	30,578

Repeat parts (a), (b), and (c) for these data.

6.12 Peterson et al. [1979] studied the patterns of infant deaths (especially SIDS) in King County, Washington during the years 1969–1977. They compared the SIDS deaths with a 1% sample of all births during the time period specified. Tables relating the occurrence of SIDS with maternal age less than or equal to 19 years of age, and to birth order greater than 1, follow for those with single births.

	Child			Child	
Birth Order	SIDS	Control	Maternal Age	SIDS	Control
$>$1	201	689	$\leq$19	76	164
$=$1	92	626	$>$19	217	1151

	Child	
	SIDS	**Control**
Birth order >1 *and* maternal age ≤ 19	26	17
Birth order $=1$ *or* maternal age >19	267	1298
Birth order >1 *and* maternal age ≤ 19	26	17
Birth order $=1$ *and* maternal age >19	42	479

(a) Compute the odds ratios and 95% confidence intervals for the data in these tables.

(b) Which pair of entries in the second table do you think best reflects the risk of both risk factors at once? Why? (There is not a definitely correct answer.)

***(c)** The control data represent a 1% sample of the population data. Knowing this, how would you estimate the relative risk directly?

6.13 Rosenberg et al. [1980] studied the relationship between coffee drinking and myocardial infarction in young women aged 30–49 years. This retrospective study included 487 cases hospitalized for the occurrence of a myocardial infarction (MI). Nine hundred eighty controls hospitalized for an acute condition (trauma, acute cholecystitis, acute respiratory diseases, and appendicitis) were selected. Data for consumption of five or more cups of coffee containing caffeine were:

Cups per Day	**MI**	**Control**
≥ 5	152	183
<5	335	797

Compute the odds ratio of a MI for heavy (≥ 5 cups per day) coffee drinkers vs. nonheavy coffee drinkers. Find the 90% confidence interval for the odds ratio.

6.14 The data of Problem 6.13 were considered to be possibly confounded with smoking. The 2×2 tables by smoking status, in cigarettes per day, are displayed in Table 6.11.

(a) Compute the Mantel–Haenszel estimate of the odds ratio and the chi-square statistic for association. Would you reject the null hypothesis of no association between coffee drinking and myocardial infarction at the 5% significance level?

(b) Using the log odds ratio as the measure of association in each table, compute the chi-square statistic for association. Find the estimated overall odds ratio and a 95% confidence interval for this quantity.

6.15 The paper of Remein and Wilkerson [1961] considers screening tests for diabetes. The Somogyi–Nelson (venous) blood test (data at 1 hour after a test meal and using 130 mg per 100 mL as the blood sugar cutoff level) gives the following table:

Test	**Diabetic**	**Nondiabetic**	**Total**
+	59	48	107
−	11	462	473
Total	70	510	580

Table 6.11 2 × 2 Tables for Problem 6.14

Cups per Day	MI	Control
Never smoked		
≥5	7	31
<5	55	269
Former smoker		
≥5	7	18
<5	20	112
1–14 cigarettes per day		
≥5	7	24
<5	33	114
15–24 cigarettes per day		
≥5	40	45
<5	88	172
25–34 cigarettes per day		
≥5	34	24
<5	50	55
35–44 cigarettes per day		
≥5	27	24
<5	55	58
45+ cigarettes per day		
≥5	30	17
<5	34	17

(a) Compute the sensitivity, specificity, predictive value of a positive test, and predictive value of a negative test.

(b) Using the sensitivity and specificity of the test as given in part (a), plot curves of the predictive values of the test vs. the percent of the population with diabetes (0 to 100%). The first curve will give the probability of diabetes given a positive test. The second curve will give the probability of diabetes given a negative test.

6.16 Remein and Wilkerson [1961] present tables showing the trade-off between sensitivity and specificity that arises by changing the cutoff value for a positive test. For blood samples collected 1 hour after a test meal, three different blood tests gave the data given in Table 6.12.

(a) Plot three curves, one for each testing method, on the same graph. Let the vertical axis be the sensitivity and the horizontal axis be (1 − specificity) of the test. The curves are generated by the changing cutoff values.

(b) Which test, if any, looks most promising? Why? (See also Note 6.7)

6.17 Data of Sartwell et al. [1969] that examine the relationship between thromboembolism and oral contraceptive use are presented below for several subsets of the population. For each subset:

(a) Perform McNemar's test for a case–control difference (5% significance level).

(b) Estimate the relative risk.

(c) Find an appropriate 90% confidence interval for the relative risk.

Table 6.12 Blood Sugar Data for Problem 6.16

Blood Sugar (mg/100 mL)	Type of Test					
	Somogyi–Nelson		Folin–Wu		Anthrone	
	SENS	SPEC	SENS	SPEC	SENS	SPEC
70	—	—	100.0	8.2	100.0	2.7
80	—	1.6	97.1	22.4	100.0	9.4
90	100.0	8.8	97.1	39.0	100.0	22.4
100	98.6	21.4	95.7	57.3	98.6	37.3
110	98.6	38.4	92.9	70.6	94.3	54.3
120	97.1	55.9	88.6	83.3	88.6	67.1
130	92.9	70.2	78.6	90.6	81.4	80.6
140	85.7	81.4	68.6	95.1	74.3	88.2
150	80.0	90.4	57.1	97.8	64.3	92.7
160	74.3	94.3	52.9	99.4	58.6	96.3
170	61.4	97.8	47.1	99.6	51.4	98.6
180	52.9	99.0	40.0	99.8	45.7	99.2
190	44.3	99.8	34.3	100.0	40.0	99.8
200	40.0	99.8	28.6	100.0	35.7	99.8

For nonwhites:

	Case	
Control	**Yes**	**No**
Yes	3	3
No	11	9

For married:

	Case	
Control	**Yes**	**No**
Yes	8	10
No	41	46

and for ages 15–29:

	Case	
Control	**Yes**	**No**
Yes	5	33
No	7	57

6.18 Janerich et al. [1980] compared oral contraceptive use among mothers of malformed infants and matched controls who gave birth to healthy children. The controls were matched for maternal age and race of the mother. For each of the following, estimate the odds ratio and form a 90% confidence interval for the odds ratio.

(a) Women who conceived while using the pill or immediately following pill use.

	Case	
Control	**Yes**	**No**
Yes	1	33
No	49	632

(b) Women who experienced at least one complete pill-free menstrual period prior to conception.

	Case	
Control	**Yes**	**No**
Yes	38	105
No	105	467

(c) Cases restricted to major structural anatomical malformations; use of oral contraceptives after the last menstrual period or in the menstrual cycle prior to conception.

	Case	
Control	**Yes**	**No**
Yes	0	21
No	45	470

(d) As in part (c) but restricted to mothers of age 30 or older.

	Case	
Control	**Yes**	**No**
Yes	0	1
No	6	103

6.19 Robinette et al. [1980] studied the effects on health of occupational exposure to microwave radiation (radar). The study looked at groups of enlisted naval personnel who were enrolled during the Korean War period. Find 95% confidence intervals for the percent of men dying of various causes, as given in the data below. Deaths were recorded that occurred during 1950–1974.

(a) Eight of 1412 aviation electronics technicians died of malignant neoplasms.

(b) Six of the 1412 aviation electronics technicians died of suicide, homicide, or other trauma.

(c) Nineteen of 10,116 radarmen died by suicide.

(d) Sixteen of 3298 fire control technicians died of malignant neoplasms.

(e) Three of 9253 radiomen died of infective and parasitic disease.

(f) None of 1412 aviation electronics technicians died of infective and parasitic disease.

6.20 The following data are also from Robinette et al. [1980]. Find 95% confidence intervals for the population percent dying based on these data: (1) 199 of 13,078 electronics technicians died of disease; (2) 100 of 13,078 electronics technicians died of circulatory disease; (3) 308 of 10,116 radarmen died (of any cause); (4) 441 of 13,078 electronics technicians died (of any cause); (5) 103 of 10,116 radarmen died of an accidental death.

(a) Use the normal approximation to the Poisson distribution (which is approximating a binomial distribution).

(b) Use the large-sample binomial confidence intervals (of Section 6.2.6). Do you think the intervals are similar to those calculated in part (a)?

6.21 Infant deaths in King County, Washington were grouped by season of the year. The number of deaths by season, for selected causes of death, are listed in Table 6.13.

Table 6.13 Death Data for Problem 6.21

	Season			
	Winter	Spring	Summer	Autumn
Asphyxia	50	48	46	34
Immaturity	30	40	36	35
Congenital malformations	95	93	88	83
Infection	40	19	40	43
Sudden infant death syndrome	78	71	87	86

(a) At the 5% significance level, test the hypothesis that SIDS deaths are uniformly ($p = 1/4$) spread among the seasons.

(b) At the 10% significance level, test the hypothesis that the deaths due to infection are uniformly spread among the seasons.

(c) What can you say about the p-value for testing that asphyxia deaths are spread uniformly among seasons? Immaturity deaths?

6.22 Fisher [1958] (after [Carver, 1927]) provided the following data on 3839 seedlings that were progeny of self-fertilized heterozygotes (each seedling can be classified as either starchy or sugary and as either green or white):

Number of Seedlings	Green	White	Total
Starchy	1997	906	2903
Surgary	904	32	936
Total	2901	938	3839

(a) On the assumption that the green and starchy genes are dominant and that the factors are independent, show that by Mendel's law that the ratio of expected frequencies (starchy green, starchy white, sugary green, sugary white) should be 9 : 3 : 3 : 1.

(b) Calculate the expected frequencies under the hypothesis that Mendel's law holds and assuming 3839 seedlings.

(c) The data are multinomial with parameters π_1, π_2, π_3, and π_4, say. What does Mendel's law imply about the relationships among the parameters?

(d) Test the goodness of fit.

6.23 Fisher [1958] presented data of Geissler [1889] on the number of male births in German families with eight offspring. One model that might be considered for these data is the binomial distribution. This problem requires a goodness-of-fit test.

(a) Estimate π, the probability that a birth is male. This is done by using the estimate $p =$ (total number of male births)/(total number of births). The data are given in Table 3.10.

(b) Using the p of part (a), find the binomial probabilities for number of boys $= 0$, 1, 2, 3, 4, 5, 6, 7, and 8. Estimate the expected number of observations in each cell if the binomial distribution is correct.

(c) Compute the X^2 value.

(d) The X^2 distribution lies between chi-square distributions with what two degrees of freedom? (Refer to Section 6.6.4)

***(e)** Test the goodness of fit by finding the two critical values of part (d). What can you say about the p-value for the goodness-of-fit test?

***6.24** **(a)** Let $R(n)$ be the number of ways to arrange n distinct objects in a row. Show that $R(n) = n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$. By definition, $R(0) = 1$. *Hint:* Clearly, $R(1) = 1$. Use *mathematical induction*. That is, show that if $R(n-1) = (n-1)!$, then $R(n) = n!$. This would show that for all positive integers n, $R(n) = n!$. Why? [To show that $R(n) = n!$, suppose that $R(n-1) = (n-1)!$. Argue that you may choose any of the n objects for the first position. For each such choice, the remaining $n-1$ objects may be arranged in $R(n-1) = (n-1)!$ different ways.]

(b) Show that the number of ways to select k objects from n objects, denoted by $\binom{n}{k}$ (the binomial coefficient), is $n!/((n-k)!\,k!)$. *Hint*: We will choose the k objects by arranging the n objects in a row; the first k objects will be the ones we select. There are $R(n)$ ways to do this. When we do this, we get the *same* k objects many times. There are $R(k)$ ways to arrange the *same* k objects in the first k positions. For each such arrangement, the other $n-k$ objects may be arranged in $R(n-k)$ ways. The number of ways to arrange these objects is $R(k)R(n-k)$. Since each of the k objects is counted $R(k)R(n-k)$ times in the $R(n)$ arrangements, the number of different ways to select k objects is

$$\frac{R(n)}{R(k)R(n-k)} = \frac{n!}{k!\,(n-k)!}$$

from part (a). Then check that

$$\binom{n}{n} = \binom{n}{0} = 1$$

(c) Consider the binomial situation: n independent trials each with probability π of success. Show that the probability of k successes

$$b(k; n, \pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

Hint: Think of the n trials as ordered. There are $\binom{n}{k}$ ways to choose the k trials that give a success. Using the independence of the trials, argue that the probability of the k trials being a success is $\pi^k(1-\pi)^{n-k}$.

(d) Compute from the definition of $b(k; n, \pi)$: (i) $b(3; 5, 0.5)$; (ii) $b(3; 3, 0.3)$; (iii) $b(2; 4, 0.2)$; (iv) $b(1; 3, 0.7)$; (v) $b(4; 6, 0.1)$.

6.25 In Section 6.2.3 we presented procedures for two-sided hypothesis tests with the binomial distribution. This problem deals with one-sided tests. We present the procedures for a test of $H_0 : \pi \geq \pi_0$ vs. $H_A : \pi < \pi_0$. [The same procedures would be used for $H_0 : \pi = \pi_0$ vs. $H_A : \pi < \pi_0$. For $H_0 : \pi \leq \pi_0$ vs. $H_A : \pi > \pi_0$, the procedure would be modified (see below).]

Procedure A: To construct a significance test of $H_0 : \pi \geq \pi_0$ vs. $H_a : \pi < \pi_0$ at significance level α:

(a) Let Y be binomial n, π_0, and $p = Y/n$. Find the largest c such that $P[p \leq c] \leq \alpha$.

(b) Compute the actual significance level of the test as $P[p \leq c]$.

(c) Observe p. Reject H_0 if $p \leq c$.

Procedure B: The p-value for the test if we observe p is $P[\widetilde{p} \leq p]$, where p is the *fixed* observed value and $\widetilde{p}$ equals $\widetilde{Y}/n$, where $\widetilde{Y}$ is binomial n, π_0.

(a) In Problem 6.2, let π be the probability of losing weight. (i) Find the critical value c for testing $H_0 : \pi \geq 1/2$ vs. $H_A : \pi < 1/2$ at the 10% significance level. (ii) Find the one-sided p-value for the data of Problem 6.2.

(b) Modify procedures A and B for the hypotheses $H_0 : \pi \leq \pi_0$ vs. $H_A : \pi > \pi_0$.

***6.26** Using the terminology and notation of Section 6.3.1, we consider proportions of success from two samples of size $n_{1\cdot}$ and $n_{2\cdot}$. Suppose that we are told that there are $n_{\cdot 1}$ total successes. That is, we observe the following:

	Success	**Failures**	
Sample 1	?		$n_{1\cdot}$
Sample 2			$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

If both populations are equally likely to have a success, what can we say about n_{11}, the number of successes in population 1, which goes in the cell with the question mark?

Show that

$$P[n_{11} = k] = \binom{n_{1\cdot}}{k} \binom{n_{2\cdot}}{n_{\cdot 1} - k} \Big/ \binom{n_{\cdot\cdot}}{n_{\cdot 1}}$$

for $k \leq n_{1\cdot}$, $k \leq n_{\cdot 1}$, and $n_{\cdot 1} - k \leq n_{2\cdot}$. *Note*: $P[n_{11} = k]$, which has the parameters $n_{1\cdot}, n_{2\cdot}$, and $n_{\cdot 1}$, is called a *hypergeometric probability*. *Hint*: As suggested in Section 6.3.1, think of each trial (in sample 1 or 2) as a ball [purple ($n_{1\cdot}$) or gold ($n_{2\cdot}$)].

Since successes are equally likely in either population, any ball is as likely as any other to be drawn in the $n_{\cdot 1}$ successes. All subsets of size $n_{\cdot 1}$ are equally likely, so the probability of k successes is the number of subsets with k purple balls divided by the total number of subsets of size $n_{\cdot 1}$. Argue that the first number is $\binom{n_{1\cdot}}{k}\binom{n_{2\cdot}}{n_{\cdot 1}-k}$ and the second is $\binom{n_{\cdot\cdot}}{n_{\cdot 1}}$.

6.27 This problem gives more practice in finding the sample sizes needed to test for a difference in two binomial populations.

(a) Use Figure 6.2 to find *approximate* two-sided sample sizes *per group* for $\alpha = 0.05$ and $\beta = 0.10$ when (i) $P_1 = 0.5, P_2 = 0.6$; (ii) $P_1 = 0.20, P_2 = 0.10$; (iii) $P_1 = 0.70, P_2 = 0.90$.

(b) For each of the following, find one-sided sample sizes *per group* as needed from the formula of Section 6.3.3. (i) $\alpha = 0.05, \beta = 0.10, P_1 = 0.25, P_2 = 0.10$; (ii) $\alpha = 0.05, \beta = 0.05, P_1 = 0.60, P_2 = 0.50$; (iii) $\alpha = 0.01, \beta = 0.01, P_1 = 0.15, P_2 = 0.05$; (iv) $\alpha = 0.01, \beta = 0.05, P_1 = 0.85, P_2 = 0.75$. To test π_1 vs. π_2, we need the same sample size as we would to test $1 - \pi_1$ vs. $1 - \pi_2$. Why?

6.28 You are examined by an excellent screening test (sensitivity and specificity of 99%) for a rare disease (0.1% or 1/1000 of the population). Unfortunately, the test is positive. What is the probability that you have the disease?

***6.29** **(a)** Derive the rule of threes defined in Note 6.9.

(b) Can you find a similar constant to set up a 99% confidence interval?

***6.30** Consider the matched pair data of Problem 6.17: What null hypothesis does the usual chi-square test for a 2×2 table test on these data? What would you decide about the matching if this chi-square was not significant (e.g., the "married" table)?

REFERENCES

Beyer, W. H. (ed.) [1968]. *CRC Handbook of Tables for Probability and Statistics*. CRC Press, Cleveland, OH.

Borer, J. S., Rosing, D. R., Miller, R. H., Stark, R. M., Kent, K. M., Bacharach, S. L., Green, M. V., Lake, C. R., Cohen, H., Holmes, D., Donahue, D., Baker, W., and Epstein, S. E. [1980]. Natural history of left ventricular function during 1 year after acute myocardial infarction: comparison with clinical, electrocardiographic and biochemical determinations. *American Journal of Cardiology*, **46**: 1–12.

Box, J. F. [1978]. *R. A. Fisher: The Life of a Scientist*. Wiley, New York.

Breslow, N. E., and Day, N. E. [1980]. *Statistical Methods in Cancer Research*, Vol. 1, *The Analysis of Case–Control Studies*, IARC Publication 32. International Agency for Research in Cancer, Lyon, France.

Bucher, K. A., Patterson, A. M., Elston, R. C., Jones, C. A., and Kirkman, H. N., Jr. [1976]. Racial difference in incidence of ABO hemolytic disease. *American Journal of Public Health*, **66**: 854–858. Copyright © 1976 by the American Public Health Association.

Carver, W. A. [1927]. A genetic study of certain chlorophyll deficiencies in maize. *Genetics*, **12**: 415–440.

Cavalli-Sforza, L. L., and Bodmer, W. F. [1999]. *The Genetics of Human Populations*. Dover Publications, New York.

Chernoff, H., and Lehmann, E. L. [1954]. The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics*, **25**: 579–586.

Comstock, G. W., and Partridge, K. B. [1972]. Church attendance and health. *Journal of Chronic Diseases*, **25**: 665–672. Used with permission of Pergamon Press, Inc.

Conover, W. J. [1974]. Some reasons for not using the Yates continuity correction on 2×2 contingency tables (with discussion). *Journal of the American Statistical Association*, **69**: 374–382.

Edwards, A. W. F., and Fraccaro, M. [1960]. Distribution and sequences of sex in a selected sample of Swedish families. *Annals of Human Genetics, London*, **24**: 245–252.

Feigl, P. [1978]. A graphical aid for determining sample size when comparing two independent proportions. *Biometrics*, **34**: 111–122.

Fisher, L. D., and Patil, K. [1974]. Matching and unrelatedness. *American Journal of Epidemiology*, **100**: 347–349.

Fisher, R. A. [1936]. Has Mendel's work been rediscovered? *Annals of Science*, **1**: 115–137.

Fisher, R. A. [1958]. *Statistical Methods for Research Workers*, 13th ed. Oliver & Boyd, London.

Fisher, R. A., Thornton, H. G., and MacKenzie, W. A. [1922]. The accuracy of the plating method of estimating the density of bacterial populations. *Annals of Applied Biology*, **9**: 325–359.

Fleiss, J. L., Levin, B., and Park, M. C. [2003]. *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, New York.

Geissler, A. [1889]. Beiträge zur Frage des Geschlechts Verhältnisses der Geborenen. *Zeitschrift des K. Sachsischen Statistischen Bureaus*.

Graunt, J. [1662]. *Natural and Political Observations Mentioned in a Following Index and Made Upon the Bills of Mortality*. Given in part in Newman, J. R. (ed.) [1956]. *The World of Mathematics*, Vol. 3. Simon & Schuster, New York.

Grizzle, J. E. [1967]. Continuity correction in the χ^2-test for 2×2 tables. *American Statistician*, **21**: 28–32.

Hanley, J. A., and Lippman-Hand, A. [1983]. If nothing goes wrong, is everything alright? *Journal of the American Medical Association*, **249**: 1743–1745.

Janerich, D. T., Piper, J. M., and Glebatis, D. M. [1980]. Oral contraceptives and birth defects. *American Journal of Epidemiology*, **112**: 73–79.

Karlowski, T. R., Chalmers, T. C., Frenkel, L. D., Zapikian, A. Z., Lewis, T. L., and Lynch, J. M. [1975]. Ascorbic acid for the common cold: a prophylactic and therapeutic trial. *Journal of the American Medical Association*, **231**: 1038–1042.

Kelsey, J. L., and Hardy, R. J. [1975]. Driving of motor vehicles as a risk factor for acute herniated lumbar intervertebral disc. *American Journal of Epidemiology*, **102**: 63–73.

Kendall, M. G., and Stuart, A. [1967]. *The Advanced Theory of Statistics*, Vol. 2, *Inference and Relationship*. Hafner, New York.

Kennedy, J. W., Kaiser, G. W., Fisher, L. D., Fritz, J. K., Myers, W., Mudd, J. G., and Ryan, T. J. [1981]. Clinical and angiographic predictors of operative mortality from the collaborative study in coronary artery surgery (CASS). *Circulation*, **63**: 793–802.

Lawson, D. H., and Jick, H. [1976]. Drug prescribing in hospitals: an international comparison. *American Journal of Public Health*, **66**: 644–648.

Little, R. J. A. [1989]. Testing the equality of two independent binomial proportions. *American Statistician*, **43**: 283–288.

Mantel, N., and Greenhouse, S. W. [1968]. What is the continuity correction? *American Statistician*, **22**: 27–30.

Mantel, N., and Haenszel, W. [1959]. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**: 719–748.

Mantel, N., Brown, C., and Byar, D. P. [1977]. Tests for homogeneity of effect in an epidemiologic investigation. *American Journal of Epidemiology*, **106**: 125–129.

Mendel, G. [1866]. Versuche über Pflanzenhybriden. *Verhandlungen Naturforschender Vereines in Brunn*, **10**: 1.

Meyer, M. B., Jonas, B. S., and Tonascia, J. A. [1976]. Perinatal events associated with maternal smoking during pregnancy. *American Journal of Epidemiology*, **103**: 464–476.

Miettinen, O. S. [1970]. Matching and design efficiency in retrospective studies. *American Journal of Epidemiology*, **91**: 111–118.

Odeh, R. E., Owen, D. B., Birnbaum, Z. W., and Fisher, L. D. [1977]. *Pocket Book of Statistical Tables*. Marcel Dekker, New York.

Ounsted, C. [1953]. The sex ratio in convulsive disorders with a note on single-sex sibships. *Journal of Neurology, Neurosurgery and Psychiatry*, **16**: 267–274.

Owen, D. B. [1962]. *Handbook of Statistical Tables*. Addison-Wesley, Reading, MA.

Peterson, D. R., van Belle, G., and Chinn, N. M. [1979]. Epidemiologic comparisons of the sudden infant death syndrome with other major components of infant mortality. *American Journal of Epidemiology*, **110**: 699–707.

Peterson, D. R., Chinn, N. M., and Fisher, L. D. [1980]. The sudden infant death syndrome: repetitions in families. *Journal of Pediatrics*, **97**: 265–267.

Pepe, M. S. [2003]. *The Statistical Evaluation of Medical Tests for Clarification and Prediction*. Oxford University Press, Oxford.

Remein, Q. R., and Wilkerson, H. L. C. [1961]. The efficiency of screening tests for diabetes. *Journal of Chronic Diseases*, **13**: 6–21. Used with permission of Pergamon Press, Inc.

Robinette, C. D., Silverman, C., and Jablon, S. [1980]. Effects upon health of occupational exposure to microwave radiation (radar). *American Journal of Epidemiology*, **112**: 39–53.

Rosenberg, L., Slone, D., Shapiro, S., Kaufman, D. W., Stolley, P. D., and Miettinen, O. S. [1980]. Coffee drinking and myocardial infarction in young women. *American Journal of Epidemiology*, **111**: 675–681.

Sartwell, P. E., Masi, A. T., Arthes, F. G., Greene, G. R., and Smith, H. E. [1969]. Thromboembolism and oral contraceptives: an epidemiologic case–control study. *American Journal of Epidemiology*, **90**: 365–380.

Schlesselman, J. J. [1982]. *Case–Control Studies: Design, Conduct, Analysis*. Monographs in Epidemiology and Biostatistics. Oxford University Press, New York.

Shapiro, S., Goldberg, J. D., and Hutchinson, G. B. [1974]. Lead time in breast cancer detection and implications for periodicity of screening. *American Journal of Epidemiology*, **100**: 357–366.

Smith, J. P., Delgado, G., and Rutledge, F. [1976]. Second-look operation in ovarian cancer. Cancer, **38**: 1438–1442. Used with permission from J. B. Lippincott Company.

Starmer, C. F., Grizzle, J. E., and Sen, P. K. [1974]. Comment. *Journal of the American Statistical Association*, **69**: 376–378.

U.S. Department of Health, Education, and Welfare [1964]. *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. U.S. Government Printing Office, Washington, DC.

von Bortkiewicz, L. [1898]. *Das Gesetz der Kleinen Zahlen*. Teubner, Leipzig.

Weber, A., Jermini, C., and Grandjean, E. [1976]. Irritating effects on man of air pollution due to cigarette smoke. *American Journal of Public Health*, **66**: 672–676.