CHAPTER 13. CONTINGENCY TABLES

When members of a given *set* can be *classified* in several different ways, a common form of statistical problem arises if a *sample* of members is drawn from the given set. For example, we might begin with the set of U.S. adult males who were alive in 1964, take one classification to be *developed lung cancer*/*did not develop lung cancer*, and take a second classification to be *heavy smoker since 1964*/*not a heavy smoker since 1964*. In such a case, we might be interested in whether or not the two classifications appear from the sample to be *associated* in some way, where *association* means that individuals in some category in one classification have a greater tendency to be in a certain category in the other classification. For example, we would say that the above two classifications appeared (from the sample) to be *associated* if individuals in the *lung cancer* category in the first classification seemed to have a greater tendency also to occur in the *heavy smoker* category in the second classification. To put it in another way, we would be asking, "Is whether or not a person develops lung cancer *associated* with whether or not the person is a heavy smoker?" (In problems of this kind, the given set from which the sample is taken is sometimes called the *population*.)

A *classification* may have two categories (for example, *has red hair*/*does not have red hair*) or it may have more than two categories. For example, the classification *scored 90 or above*/*scored below 90 but above 80*/*scored 80 or below* has three categories. In a 2-category classification, we sometimes choose one of the classes and call it the *attribute* or *characteristic* of the classification. Thus in the example above, we might speak of the attribute *heavy smoker since 1964* and of the attribute *developed lung cancer*.

In what follows, we shall be largely concerned with cases where there are two classifications, and for the most part, we shall only consider cases where each of the two classifications is 2-category. When we say that two *attributes* are *associated*, we shall allow the association to be either positive or negative. (For example, we may find a negative association between the attribute *heavy smoking* and the attribute *freedom from cancer*.)

Tables. When there are two attributes in question (when we have two classifications, and each is 2-category), the observed data can be presented in a single 2 x 2 table. For example, we might have the following data on a random sample of 10,000 U.S. males who were living in 1964. (The specific data are fictitious.)

	Cancer	Non-Cancer
Heavy smoker	50	950
Not heavy smoker	10	8990

Such a table is called a contingency table. In working with a contingency table, we begin by forming the sum of each row (these sums are called the right-hand margin) and the sum of each column (these sums are called the bottom margin). Each margin has, as its sum, the total number of individuals observed. If we write in margins for the above table, we get

50	950	1,000
10	8,990	9,000
60	9,940	10,000

The basic statistical question in considering a contingency table from a given sample is the following: if the table by itself appears to suggest a possible association, can this apparent association be reasonably explained by asserting that there is no association in the underlying population and that what we see in the table is a random fluctuation (in the sampling process), or must we conclude that the observed fluctuation would be highly unlikely and that some amount of association really exists in the underlying population? In analyzing a contingency table, our approach to this question will be the following. First, we shall assume that no association exists. Second, we shall then calculate a DLS for the given observation (that is to say, for the given table) on the assumption that no association exists.

If the DLS is small, this will suggest to us that some association does in fact exist.

To find a *DLS*, we must assume a model and a metric, or else a composite model and a well-defined metric; and this model (or composite model) must represent *non-association*. We shall now see that there are three basically different types of experimental procedure for drawing a sample from a population and forming a contingency table. Corresponding to these three types of procedure, we shall see that there are three different kinds of model for non-association. Ultimately, however, we shall find that the same calculations for finding the *DLS* are used for each of the three types. We initially limit our consideration to 2x2 tables.

Type α: *Multinomial procedures*; *multinomial models for non-association*. Let us call the two attributes A and B. Let A and $\bar{A}$ be the two categories of one classification, and let B and $\bar{B}$ be the two categories of the other classification.

Experimental procedure: Individuals are drawn at random (with replacement) from the given population without regard to the attributes they possess, and for each individual the occurrence or non-occurrence of A and B is recorded.

Models for the procedure: The four possible combinations of attributes for an individual must occur with certain probabilities $P(A \,\&\, B)$, $P(A \,\&\, \bar{B})$, $P(\bar{A} \,\&\, B)$, and $P(\bar{A} \,\&\, \bar{B})$. Any set of values p_1, p_2, p_3, p_4 for these probabilities, subject to the condition that $p_1+p_2+p_3+p_4 = 1$, constitutes a model for the experimental procedure.

The probability of an observed table will be given by the multinomial formula with these probabilities. For example, if $P(A \& B) = 0.1$, $P(A \& \overline{B}) = 0.2$, $P(\overline{A} \& B) = 0.3$, and $P(\overline{A} \& \overline{B}) = 0.4$, then the particular observation

	B	$\overline{B}$
A	a	b
$\overline{A}$	c	d

will occur with probability

$$\frac{n!}{a!b!c!d!}(0.1)^a(0.2)^b(0.3)^c(0.4)^d ,$$

where $n = a + b + c + d$.

Remark. In practice, the sample will usually be drawn without replacement. It is almost always the case that the underlying population is so large that the difference, in theoretical results, between drawing with replacement and drawing without replacement is negligible. ~~(see the exercises at the end of this chapter.)~~ It is easier and simpler, in theoretical calculations, to assume that drawing occurs with replacement. Hence we shall make this assumption. This remark applies to procedures of type β and type γ (to be described below) as well as to procedures of type α.

Models for non-association: For which models, under this procedure, can we say that A and B are not associated? We take non-association to mean that the events A occurs and B occurs (for an individual) are independent (that is to say,

$P(A \& B) = p_A p_B$, where $p_A = P(A) = P(A \& B) + P(A \& \bar{B})$, and $p_B = P(B) = P(A \& B) + P(\bar{A} \& B))$. (In the example above, $p_A = 0.1 + 0.2 = 0.3$ and $p_B = 0.1 + 0.3 = 0.4$, so that $P(A \& B) = 0.1 \neq 0.12 = p_A p_B$, ana independence does not hold.) If independence holds, and hence the attributes are non-associated, we can calculate the probability of each of the four possible combinations of attributes from the two values p_A and p_B. For example, if we know that $p_A = 0.3$ and $p_B = 0.4$, and if we assume independence, we have

$$\begin{aligned} P(A \& B) &= (0.3)(0.4) = 0.12, \\ P(A \& \bar{B}) &= (0.3)(0.6) = 0.18, \\ P(\bar{A} \& B) &= (0.7)(0.4) = 0.28, \\ P(\bar{A} \& \bar{B}) &= (0.7)(0.6) = 0.42. \end{aligned}$$

It follows that, for this model of non-association, the occurrence of the particular observation

$$\begin{array}{c|c|c|l} & B & \bar{B} & \\ \hline A & a & b & \\ \hline \bar{A} & c & d & \\ \hline & & & n \end{array}$$

will have a probability given by the multinomial formula

$$\frac{n!}{a!b!c!d!} (0.12)^a (0.18)^b (0.28)^c (0.42)^d$$

where $n = a+b+c+d$. More generally, given a model (p_A, p_B) for non-association (for a type α procedure), the probability of a particular table can be written:

$$P\left(\begin{array}{c|c} a & b \\ \hline c & d \end{array}\right) = \frac{n!}{a!b!c!d!}\ (P(A\ \&\ B))^{a}(P(A\ \&\ \overline{B}))^{b}(P(\overline{A}\ \&\ B))^{c}(P(\overline{A}\ \&\ \overline{B}))^{d}$$

$$= \frac{n!}{a!b!c!d!}\ p_A^{a+b}(1 - p_A)^{c+d}\ (p_B)^{a+c}\ (1 - p_B)^{b+d}\ ,$$

Note that the models for non-association form a composite model within the larger universe of all models for our type α procedure.

Type β: Imposing one margin:double binomial procedures; double binomial models for non-association.

Experimental procedure: Instead of drawing n individuals at random without regard to attributes A and B, we can specify that we will draw n_1 individuals at random with replacement from among those with attribute A, and n_2 individuals at random from among those who do not have A, where the numbers n_1 and n_2 are fixed ahead of time. We do this and observe, in each of the two samples, the number of individuals with attribute B. Our observation is presented in the form of a contingency table with n_1 and n_2 as the right-hand margin.

	B	$\overline{B}$	
A	a	b	n_1
$\overline{A}$	c	d	n_2

Models for the procedure. The probability of an observed table will be given by the double binomial formula. For example, if $p_1 = P(B|A) = 0.2$ is the probability for B among individuals with A, and $p_2 = P(B|\overline{A}) = 0.4$ is the probability for B among individuals with $\overline{A}$, then the above observation will occur with probability

$$\binom{n_1}{a}(0.2)^{a}(0.8)^{b}\ \binom{n_2}{c}\ (0.4)^{c}(0.6)^{d}.$$

Note that the margin (n_1, n_2) is *imposed* ahead of time by our experimental procedure. In the smoking and cancer case, this form of model would correspond to an experimental procedure of taking $n_1 = 1000$ heavy smokers and $n_2 = 9000$ individuals who are not heavy smokers, and seeing how many in each group have developed cancer. Alternatively, this form of model would also hold for a procedure of taking $n_1 = 60$ individuals who have developed cancer and $n_2 = 9940$ who have not, and seeing how many in each group are heavy smokers. (We would then take *cancer* to be A and *heavy smoker* to be B.)

Models for non-association: We now take *non-association* to mean that the probability of getting B is the same for an individual with A as for an individual with $\overline{A}$, that is, to mean that $p_1 = P(B|A) = p_2 = P(B|\overline{A})$. If we assume that this non-association holds, we can calculate the probability of the occurrence of a particular table from $p_B = p_1 = p_2$. For example, if the common probability for B for each of the two samples is assumed to be 0.3, the probability of the observation $\begin{array}{c|c} a & b \\ \hline c & d \end{array} \begin{array}{l} : n_1 \\ : n_2 \end{array}$ will be given by the double binomial formula

$$\binom{n_1}{a}(0.3)^a(0.7)^b \binom{n_2}{c}(0.3)^c(0.7)^d$$

More generally, given a model p_B for non-association in a type β procedure, the probability of a particular table can be written:

$$\binom{n_1}{a}\binom{n_2}{c} \; p_B^{a+c}(1 - p_B)^{b+d},$$

Note that the models for *non-association* form a composite model within the larger universe of all models for our type β procedure. This composite model is the set of all models with $p_1 = p_2$.

Example C in Chapter 12 was a double binomial experiment. The observation $X_1 = 0$, $X_2 = 3$ in that experiment can be given as the contingency table

	B	$\bar{B}$	
A	0	10	10
$\bar{A}$	3	7	10

where attribute A is *getting standard diet*, $\bar{A}$ is *getting supplemented diet*, attribute B is *showing signs of infection*, and $\bar{B}$ is *not showing signs of infection*. The models for non-association for this type β experiment form the composite model that was denoted M_o in Chapter 12.

Type γ: *Imposing both margins : hypergeometric procedures; the hypergeometric formula for non-association*.

Experimental procedure: In this case, as with type β, we draw n_1 individuals with A and n_2 with $\bar{A}$ (with replacement), where n_1 and n_2 are fixed ahead of time. Then, however, our experimental procedure is to divide the entire sample of $n_1+n_2 = n$ individuals into two groups of sizes m_1 and m_2, where m_1 and m_2 are fixed ahead of time and $m_1+m_2 = n$, and where the group of size m_1 is made up of those m_1 individuals who have attribute B *most strongly* . (This procedure assumes that we can compare any two individuals as to the *degree* to which they have attribute B.) Finally, in our contingency table, we take *being a member of this group of size m_1* as a new (and redefined) attribute B. Our observation is presented in the form of a contingency table with n_1 and n_2 as right-hand margin and with m_1 and m_2 as bottom margin.

a	b	n_1
c	d	n_2
m_1	m_2	

Models for the procedure. We let Y_1 be a continuous random variable which measures the *degree* to which an individual has attribute B when that individual is drawn at random from the population of all individuals with attribute A. Similarly, let Y_2 measure the degree to which an individual has attribute B when drawn at random from the population of all individuals with attribute $\bar{A}$. A *model* is a pair of functions (g_1, g_2) where g_1 is a probability density for Y_1, and g_2 is a probability density for Y_2. Given a model (g_1, g_2) what is the probability of observing the above table? The result is more complex than in the type α and type β cases, and is obtained in an exercise below.

Models for non-association. We now take non-association to mean that the probability distribution for Y_1 is the same as for Y_2. Hence a model for non-association is a pair (g_1, g_2) of probability densities where g_1 is identical with g_2. Note that the models for non-association form a composite model within the larger universe of all models for our type γ procedure. Consider the table

$$\begin{array}{c|c:c} a & b & n_1 \\ \hline c & d & n_2 \\ \hdashline m_1 & m_2 & \end{array}$$

In the special case of a model for <u>non-association</u> (a model(g_1, g_2) with $g_1 = g_2$) the probability of observing this table can be obtained by the following argument. We assume that the n_1 individuals with A are drawn first, and we identify them, in the order drawn, with the integers: $1, 2, \ldots, n_1$. The n_2 individuals with $\bar{A}$ are then drawn, and we identify them, in the order drawn, with the integers: $n_1+1, n_1+2, \ldots, n_1+n_2$. Because the distribution for Y_1 in the A group is identical with the distribution for Y_2 in the $\bar{A}$ group, and because $n_1+n_2 = n$, it follows that if the experimental procedure for forming the table were repeated many times, we would expect each subset of size m_1 from $\{1, 2, \ldots, n_1, n_1+1, \ldots, n_1+n_2\}$ to occur about equally often as the set of positions for the observed m_1 group. The total number of ways of choosing a set of positions for the m_1 group is $\binom{n_1 + n_2}{m_1} = \binom{n}{m_1}$. How many of these ways yield a individuals in the A & B cell of the table and c individuals in the $\bar{A}$ & B cell? Since there are $\binom{n_1}{a}$ ways to get the individuals for the first cell, and $\binom{n_2}{c}$ ways to get the individuals for the second, the answer is $\binom{n_1}{a}\binom{n_2}{c}$. Hence the probability of getting the observed table, if we assume non-association, is

$$\frac{\binom{n_1}{a}\binom{n_2}{c}}{\binom{n}{m_1}} \quad \text{(which can also be written} \quad \frac{\binom{n_1}{a}\binom{n-n_1}{n_1-a}}{\binom{n}{m_1}}\text{.)}$$

This formula is called the hypergeometric formula, and we abbreviate it as

$$h(a; n, n_1, m_1) .$$

It is easy to show that $h(a; n, n_1, m_1) = h(a; n, m_1, n_1)$. (See the Exercises.)

Thus, for example, if we assume some model for non-association, then the probability of getting the table $\begin{array}{c|c} 0 & 10 \\ \hline 3 & 7 \end{array}$ must be

$$\frac{\binom{10}{0}\binom{10}{3}}{\binom{20}{3}} = \frac{\frac{10\cdot 9\cdot 8}{3\cdot 2\cdot 1}}{\frac{20\cdot 19\cdot 18}{3\cdot 2\cdot 1}} = \frac{2}{19}$$

Note, as a special feature of non-association in a type γ procedure, that this probability does not depend upon the particular model of non-association that we use. It only depends upon the values occurring in the observed table.

Note also that the margin (n_1, n_2) and the margin (m_1, m_2) are both imposed ahead of time by our experimental procedure. In the smoking and cancer case, this form of model would correspond, for example, to a procedure of deciding on the values n_1, n_2, m_1 and m_2 ahead of time, then randomly taking n_1 individuals who have developed cancer and n_2 who have not, and finally dividing the entire sample into two groups where the first group is made up of those m_1 individuals in the entire sample who are the m_1 heaviest smokers. (Thus, in our final table, attribute A is developing cancer and attribute B is being one of the m_1 heaviest smokers.)

Finding the DLS. In order to get a DLS, we shall proceed as follows. In the type α case, we shall consider the set of models for non-association as a single composite model, and we shall look for a well-defined metric. In the type β case, we shall do the same. In the type γ case, we have seen that the probability of a table does not depend on the particular model used for non-association. Hence every metric will be well-defined on the composite model for non-association. In the type γ case, we need only find some intuitively suitable metric. In the type α case, we shall find a well-defined metric by a conceptual procedure much like that described in Chapter 11 (and discussed further in Chapter 12) for the composite model of all Poisson distributions. In particular, we shall use the observed table itself to indicate a particular model in the composite model, and we shall then use the CS-metric to give a distance-value for how far the observed table is from the TER of the indicated model. In the type β case, we shall proceed similarly. In both cases, the metric so obtained will prove to be well-defined on the composite model provided that the entries in the TER table are sufficiently large. In the type γ case, we shall apply the CS-metric after first identifying an appropriate TER. Finally, in each of the three cases, we shall calculate a DLS by methods to be described below. We now give all these procedures in more detail.

Using the observation to indicate the model in type α and type β. In order to get a specific model of non-association for the multinomial case (type α), two values must be given: p_A and p_B. The estimate is obvious and simple, we take p_A and p_B to be the observed relative frequencies:

$$p_A = \frac{n_1}{n}$$

$$p_B = \frac{m_1}{n} \quad :$$

(We shall see in an exercise that this is in fact the maximum-likelihood estimate.)

In the double binomial case (type β), we must give a value for $p_B = p_1 = p_2$. If we use the observation to estimate this value, we have:

$$p_B = \frac{m_1}{n} \quad .$$

(We saw in Chapter 12 (Exercise 12-2) that this is the maximum-likelihood estimate.)

In the hypergeometric case (type γ) we have seen that the probability of a given table is independent of the particular model for non-association used. Hence no estimate is needed.

<u>Finding the theoretically expected result (for the purpose of applying the CS-metric)</u>. We look at the three types. Recall from Chapter 10 that a metric for a given model measures how far an observation is from giving strongest confirmation of that model. We saw that one way to construct a metric is to identify an imaginary observation that we can view as giving strongest confirmation (we called it the <u>TER</u> (theoretically expected result)) and then to find a formula which expresses a distance between the <u>TER</u> and the observation actually obtained.

For a type α model of non-association, the probability for an individual to be in the upper-left cell is $p_A p_B$. Hence the expected number of individuals in that cell is $p_A p_B n$. Hence, for our estimated model, the expected number of individuals in this cell is

$$p_A p_B n = \frac{n_1}{n} \cdot \frac{m_1}{n} \cdot n = \frac{n_1 m_1}{n} \quad .$$

Similarly for the other three cells, and we have as our full TER:

$\frac{n_1 m_1}{n}$	$\frac{n_1 m_2}{n}$
$\frac{n_2 m_1}{n}$	$\frac{n_2 m_2}{n}$

For a type β model of non-association, the expected number for the upper-left cell is $p_B \cdot n_1$. Hence for our estimated model, the expected number of individuals in this cell is

$$p_B \cdot n = \frac{m_1}{n} \cdot n_1 = \frac{n_1 m_1}{n} \quad .$$

Similarly for the other cells, and we find that our TER is exactly the same as for type α.

For a type γ model of non-association, we shall show in an exercise that the expected number in the upper-left cell is again

$$\frac{n_1 m_1}{n} \quad .$$

Similarly for the other cells, and we find that our TER is the same as for type α and type β. We use the notation:

E_1	E_2
E_3	E_4

for the *TER* which is common to all three procedures. Thus the TER for each cell is obtained by *multiplying the marginal values corresponding to that cell and then dividing by n*. In the smoking/ cancer example, this gives, for the upper left cell,

$$E_1 = \frac{1000 \times 60}{10,000} = 6.$$

Similarly for the other cells, and we have our *TER*

6	994	1000
54	8946	9000
60	9940	

It is easy to verify that the margins for the *TER* will always be the same as for the originally observed table.

Applying the CS-metric. In all three types, we use the *CS*-metric to measure distance of our observation from the *TER*. The general formula will be:

$$\chi_o^2 = \frac{(a-E_1)^2}{E_1} + \frac{(b-E_2)^2}{E_2} + \frac{(c-E_3)^2}{E_3} + \frac{(d-E_4)^2}{E_4}$$

In the smoking/cancer example, this gives us

$$\chi_o^2 = \frac{44^2}{6} + \frac{44^2}{54} + \frac{44^2}{994} + \frac{44^2}{8946}$$

$$= 44^2\left(\frac{1}{6} + \frac{1}{54} + \frac{1}{994} + \frac{1}{8946}\right) \doteq 360.7$$

Note that in this case $|a-E_1| = |b-E_2| = |c-E_3| = |d-E_4|$. For 2 x 2 tables, this is *always* true. For example, in the general case,

$$a-E_1 = a - \frac{n_1 m_1}{n} = a - \frac{(a+b)(a+c)}{a+b+c+d} = ad-bc$$

and

$$b-E_2 = b - \frac{n_1 m_2}{n} = b - \frac{(a+b)(b+d)}{a+b+c+d} = bc-ad \quad .$$

Hence we can write, as a general formula:

$$\chi_o^2 = (a-E_1)^2\left(\frac{1}{E_1} + \frac{1}{E_2} + \frac{1}{E_3} + \frac{1}{E_4}\right)$$

We have obtained, for any given observed table ω, a corresponding value χ_o^2. We observe that the value of χ_o^2 is the same for all three types. We now define a metric as follows. For any given table $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$, we take the value of the metric to be the value χ_o^2 for that table. In the Appendix to this chapter, we shall show that for both type α and type β, the value of this metric (for a given observation) is well-defined (to 2 decimal places) on the composite model for non-association provided that the cell values of the *TER* (from the given observation) are all $\geq$ 5. It remains to describe here how the *DLS* values of an observation can be found. There are two different procedures that may be used. We call them the *large sample method* and the *small sample method*. We first describe these methods as calculations and state their chief properties. That the methods are correct, and that they indeed have the stated properties, will be shown in the Appendix.

Calculating the DLS: the large sample method. Consider a 2 x 2 table obtained under a multinomial procedure (type α) with all four entries ≥ 5. We shall calculate the DLS from the observed χ_o^2 value by using the chi-square curve with one degree of freedom. The use of one degree of freedom can be made plausible by the following argument. Since the multinomial experiment has 4 categories, since we have used two independent algebraic facts from the observation (namely, the values $p_A = \frac{n_1}{n}$ and $p_B = \frac{m_1}{n}$) to fix the model, and since we have one algebraic constraint on the observation (a+b+c+d = n), the number of degrees of freedom must be 4 - 2 - 1 = 1. For example, if we assume that the observation in the smoking/cancer example was obtained under a type α procedure, we use the value 360.7 (obtained above) in the chi-square table for $d = 1$. In this table, the chi-square value for 0.001 is 10.83, and so we conclude that the DLS is very much smaller than 0.001.

Consider next a 2x2 table obtained under a double binomial procedure (type β) with all entries of the TER table ≥ 5. We shall again calculate the DLS from the observed χ_o^2 value by using the chi-square curve with one degree of freedom. Here the use of one degree of freedom can be made plausible by the following argument. Even though we have used only one algebraic fact $(p_B = \frac{m_1}{n})$ from the observation, the fact that one margin is fixed (that is, that the experiment is double binomial rather than multinomial) places an additional algebraic constraint on the observation ($a + b = n_1$). Hence the number of degrees of freedom must be $d = 4 - 1 - 2 = 1$.

Thus the two procedures (for type α and type β) give exactly the same calculations and result in exactly the same DLS value.

Consider next a 2 x 2 table obtained under a hypergeometric procedure (type γ) with all entries of the TER table ≥ 5. In the Appendix below, we shall show that chi-square approximation can be used in this case also. In fact the approximation procedure, as a calculation, is exactly the same as in the two previous cases, with the number of degrees $d = 1$.

The continuity correction. Recall that in the theory of normal approximation, the term "$\frac{1}{2}$" appeared in the formula

$$\alpha = \frac{x \pm \frac{1}{2} - np}{\sqrt{npq}}$$

to take account of the width of bars at the ends of a bar graph whose area is being approximated by the normal curve. In the application of chi-square approximation to 2 x 2 contingency tables of type γ, a similar matter of bar-width arises, and a corresponding correction is required. The correction takes the form of inserting the term "$\frac{1}{2}$" as follows:

$$x_o^2 = (|a-E_1|-\tfrac{1}{2})^2\left(\frac{1}{E_1} + \frac{1}{E_2} + \frac{1}{E_3} + \frac{1}{E_4}\right).$$

(If $|a - E_1| \leq \frac{1}{2}$, χ_o^2 should be taken to be 0.) The correction is called the continuity correction or the Yates correction. It is only used when a 2 x 2 table is approximated by the chi-square curve with $d = 1$. We shall see in the Appendix that the correction should also be used when we do chi-square approximation for a 2 x 2 contingency table of type α or type β. For large values of χ_o^2, the correction does not significantly affect the resulting DLS. (In the smoking/cancer example, it changes the value of χ_o^2 from 360.7 to 352.6, and the DLS remains approximately 0.)

Thus the formal chi-square approximation calculation of the DLS is the same for all three types. We speak of this calculation as the large sample method.

Remark. What if, in the case of type α, the values of p_A and p_B were given ahead of time and not obtained from the model. How would we get the DLS? The discussion in Chapter 11 suggests that we use chi-square approximation with $d = 4 - 1 = 3$. What if, in the case of type β, the value of p_B were given ahead of time and not obtained from the model. How would we get the DLS? The discussion in Chapter 11 suggests that we use chi-square approximation with $d = 2$. Both suggestions can be shown to be correct.

Calculating the DLS: small sample method for type γ. When one or more of the cells of the TER table has its entry < 5, the chi-square approximation may no longer give two-decimal place accuracy. Indeed in the case of type α and type β tables, there may be no useful metric that is well-defined with respect to the composite models for non-association. Hence the DLS may not be well-defined to two decimal places. In the case of type γ, however, every metric is well-defined and the DLS can be calculated directly by getting the probability of each possible observation that is no closer to the TER than the actual observation, and by then summing these probabilities. We call this calculation the small sample method. An example is given in the illustration below.

Illustration. We assume a type γ procedure. We now do both a large sample calculation and a small sample calculation for the table

29	24
5	10

,

We begin with the large sample calculation. We first add margins, and then get the TER. We have

29	24	53
5	10	15
34	34	68

and

26.5	26.5
7.5	7.5

as the TER.

Calculating χ_o^2 , we have

$$\chi_o^2 = (|29-26.5|-\tfrac{1}{2})^2(\frac{1}{7.5} + \frac{1}{7.5} + \frac{1}{26.5} + \frac{1}{26.5})$$
$$= 1.37$$

Using the chi-square table on page 350 for d = 1, and interpolating, we get DLS = 0.25.

(To use a normal table to evaluate C_1 (see page 348), we would take

$$C_1(1.37) = 1-2A(\sqrt{1.37}) = 1-2A(1.17) = 1-2(0.3790) = 0.24$$

This is the more accurate value. With 2 x 2 contingency tables, the normal table on pages 165-166 gives slightly greater accuracy, since less interpolation is required.)

If we omitted the continuity correction, we would have

$$\chi_o^2 = (29 - 26.5)^2(\frac{1}{7.5} + \frac{1}{7.5} + \frac{1}{26.5} + \frac{1}{26.5})$$
$$= 2.14 .$$

Using the chi-square table for d = 1 and interpolating, we would then get (incorrectly) DLS = 0.15.

We next do the small sample calculation. We need only list those possible observations which have the same imposed margins as the actual observation and which are no closer to the TER than the actual observation. We measure distance by the CS-metric, or we can use, in this case, the equivalent and simpler metric $|a-E_1|$. For each observation, we calculate its probability by the hypergeometric formula:

$$P\left(\begin{array}{c|c} a & b \\ \hline c & d \end{array}\right) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = h(a; a+b+c+d, a+b, a+c)$$

For $|a - E_1| = 2.5$, we have the tables $\begin{array}{c|c} 29 & 24 \\ \hline 5 & 10 \end{array}$ and $\begin{array}{c|c} 24 & 29 \\ \hline 10 & 5 \end{array}$. Their probabilities are $h(29; 68, 53, 34) = 0.08$ and $h(24; 68, 53, 34) = 0.08$. When $|a-E_1| = 3.5$, we have the tables $\begin{array}{c|c} 30 & 23 \\ \hline 4 & 11 \end{array}$ and $\begin{array}{c|c} 23 & 30 \\ \hline 11 & 4 \end{array}$. Each of these has probability 0.03. When $|a-E_1| = 4.5$, we have $\begin{array}{c|c} 31 & 22 \\ \hline 3 & 12 \end{array}$ and $\begin{array}{c|c} 22 & 31 \\ \hline 12 & 3 \end{array}$. Each has probability 0.01. The remaining tables, for $|a-E_1| \geq 5.5$, all have probability $= 0.00$ (to two decimal places).

Summing these probabilities, we have DLS = 0.24. This result, by the small sample method, agrees (as it should) with the large sample result.

Small sample calculation for type α and type β. Let M_1 be the composite model for non-association in a type α case and let M_o be the composite model for non-association in a type β case: If the TER value for one of the cells of a given table is < 5, the DLS may not be well-defined for the composite model M_1 (in type α) or for the composite model M_o (in type β). Let $\delta(\omega)$ be the DLS value obtained by treating a given observed table ω as if it were of type γ and then using the small sample method. Using the final theorem from the Appendix to Chapter 12, we shall see in the Appendix to this chapter that: (i) if the procedure is of type α then the DLS of ω for each individual model in M_1 is nearly as small as, or smaller than, $\delta(\omega)$; and (ii) if the procedure is of type β, then the DLS of ω for each individual model in M_0 is nearly as small as, or smaller than, $\delta(\omega)$. Thus in the type α and type β cases, a small value of $\delta(\omega)$ tells us that the DLS values for all models of non-association are small. Hence a small value for $\delta(\omega)$ will cast doubt on the assumption of non-association. In this way, the small sample method can be used, for procedures of type α and type β.

r × s tables. So far, we have considered 2 × 2 tables, which arise when we have two classifications, each of them 2-category. If one classification is r-category and one is s-category, we obtain an r × s table. The theory and procedures for r × s tables are closely similar to the case of 2 × 2 tables. We briefly describe the similarities and differences and give an example.

(a) Non-association. The assumption that non-association holds is similar to before. It states that position in one classification is independent of position in the other classification.

(b) Procedures. The same three types of procedure occur as before. In type α, we have a multinomial experiment with rs categories. In type β (imposing one margin) we have r independent multinomial experiments, each of s categories.

In type γ (imposing both margins) we assume that the second classification is based on an attribute whose strength can be measured on some numerical scale and that this scale is used for dividing the entire sample into s categories of pre-assigned size and ascending strength.

(c) Models. For each type, we take an appropriate universe of models analogous to the universe used for 2 x 2 procedures.

(d) Composite models. In each type, a composite model for non-association can be defined in a form exactly analogous to the composite model previously described for 2 x 2 tables. In the case of type γ, we find, as before, that a single probability formula (for the probability of a given table) holds for all models in the composite model. This probability formula is a generalization of the hypergeometric formula.

(e) Metrics. As before, we use a CS-metric for all three types. The TER indicated by a given observation is obtained for all three types by the same general rule as for the 2 x 2 case: for each cell, multiply the two corresponding marginal values and then divide by n (where n is the total count). When $r > 2$ or $s > 2$, it is no longer true that $|a - E_1| = |b - E_2| = \ldots$. Hence the full formula for the CS-metric must be used.

(f) Large sample calculation. This is done as before, except that the number of degrees of freedom used is $d = (r-1)(s-1)$. (See the Appendix.) There is no continuity correction if either r or s is > 2. When each $E_i \geq 5$, the DLS value is well-defined (to two decimal places) over all models of non-association.

(g) Small sample calculation. A form of hypergeometric formula can be used, but the calculation may be lengthy and require a computer or programmable calculator. As before, the computation gives an exact DLS value for type γ and an upper bound to the DLS values (over all models of non-association) for types α and β.

Example. Two strains of rats are to be compared as to blood type. 32 rats of strain I and 36 rats of strain II are tested. The rats are later classified as blood type A/blood type B/blood type C. The results yield the following 2 x 3 contingency table

	A	B	C	
I	5	16	11	32
II	4	11	21	36
	9	27	32	68

where we have added the margins. The TER for the upper left cell is $E_1 = \frac{9(32)}{68} = 4.24$. Similarly for the other cells, and we get

4.24	12.71	15.06
4.76	14.29	16.94

as TER.

Calculating x_o^2, we get

$$x_o^2 = \frac{(0.76)^2}{4.24} + \frac{(3.29)^2}{12.71} + \frac{(4.06)^2}{15.06} + \frac{(0.76)^2}{4.76} + \frac{(3.29)^2}{14.29} + \frac{(4.06)^2}{16.94}$$

$$= 3.93$$

Using chi-square approximation with $d = (3-1)(2-1) = 2$, we get DLS = 0.15

In a case like the above, where two populations are being compared with respect to some classification, we sometimes express our basic assumption of non-association by saying that the two populations are *homogeneous* with respect to the given classification. In the above example, if we *assume* that the two strains *are* homogeneous, then we can *conclude* from the *DLS* that an observation as non-homogeneous as the one obtained will only occur about 15% of the time.

All the tables above are based on *two* classifications. Such tables are called *two-way* contingency tables. If three classifications (for example, smoking/cancer/age) were used, we would have a *three-way* table. A three-way table is most naturally displayed in three-dimensional form. The theory and procedure are similar to before, except that the calculation of degrees of freedom must take appropriate account of both the number of facts from the observation used to indicate a model and of the number of marginal constraints imposed. Thus, in a 2 x 2 x 2 table under type α, we have 8 cells and use 3 facts from the observation. Hence, for chi-square approximation, we would use $d = 8 - 3 - 1 = 4$. We consider this further in the Appendix.

Remark. Analysis of an observed contigency table shows us how likely we are to get an observation with this much apparent association, when we *assume* non-association. It is important to keep in mind what the analysis, as given above, does *not* show us. It does not indicate an *amount* of association that may exist, nor does it indicate the *mathematical nature* or *form* of this association. (We return to this in Chapter 15 and Chapter 17.)

We must also be cautious in interpreting a contingency table with a very small DLS. In the 2 x 2 case, a small DLS tells us that an association exists. It does not, however, give us any information as to a causal relationship between the two attributes. Either might cause the other, or indeed, both might be caused by some third (and not obvious) attribute. Further detailed scientific study is often required to settle questions of this kind. Before such study we could not be sure, for example, whether, on the one hand, smoking caused cancer or, on the other hand, both were caused by some third environmental or genetic factor. (In fact, in the case of smoking and cancer, such further study has occurred. Study of tissue changes in animals exposed to tobacco smoke indicates that a direct causal relationship does indeed exist.)

EXERCISES FOR CHAPTER 13

13-1. In a sample of 1000 Frenchmen, 59 percent own television and of these 29 percent buy books. Of those without television 22 percent buy books.

(a) Which type of procedure appears to have been followed to obtain this data.

(b) Find an approximate DLS under the assumption of no association.

13-2. An anthropologist studying ethnic types in the British Isles obtained the following observations in a certain district:

19 men had a'long' head and red hair;
46 men had a 'long' head but did not have red hair;
8 men had a 'short' head and red hair.
73 men had a 'short' head but did not have red hair.

(a) Which type of procedure appears to have been followed to obtain this data.

(b) Find an approximate DLS under the assumption of no association.

13-3. In an investigation of how women held a baby against them as compared with how they held an inanimate object, the following counts were obtained:

When handed a baby, 49 women placed it first against the left side of their chest and 7 placed it against the right. When handed a large parcel 33 women placed it first against the left side of their chest and 31 placed it against the right.

(a) Which type of procedure appears to have been followed to obtain this data.

(b) Find an approximate DLS under the assumption of no association.

13-4. A random sample of 1000 males between 16 and 24 years old is taken. The sample is divided into younger and older halves. The sample is divided into the quarter (250) who have been least employed during the past six months and the three quarters who have been more employed. 200 of the least employed are in the younger half.

You may assume that the same probability formula (assuming no association) holds for this procedure as for a type γ procedure. (See 13-5.) Use the large sample method to estimate the DLS of this observation on the assumption of no association.

13-5. Consider the observation procedure described in 13-4.

(a) Define a set of possible models for this procedure.

(b) Define the subset of models which represent non-association.

(c) Show that the probability of an observed table, assuming non-association, is given by the hypergeometric formula. (Hint: The proof is similar to the proof on page 416.)

13-6. The following contingency table is obtained by a procedure in which both margins are fixed. Do both a large sample and a small sample calculation of the DLS for this table, assuming no association.

10	5
5	10

13-7. Under the assumption of no association, find an upper bound on the DLS for a double binomial experiment of 10 trials and 10 trials in which the observation (0,3) is obtained.

13-8. Two vaccines for mumps (A,B) were compared with a placebo in a clinical trial; the numbers of children uninfected, mildly infected and severly infected in the following 24-month period were as follows:

	Placebo	A	B
Uninfected	100	146	149
Mild	71	32	28
Severe	29	17	16

Find a DLS for this observation, assuming that neither vaccine acts any differently than the placebo.

13-9. In 13-8, consider only the two vaccines. Find a DLS assuming that both vaccines act in the same way.

13-10. Show that $h(a;\ n,\ n_1,\ m_1) = h(a;\ n,\ m_1,\ n_1)$.

13-11. A pharmaceutical company has tested a new food supplement for the possibility that it may prevent colds. The test was carried out as follows. 14 individuals were selected initially. 7 of these 14 were then randomly selected to form an experimental group while the remaining 7 served as a control group. The experimental group were given the supplement for one year, and the control group were not. During the course of the year, one member of the experimental group had a cold and 6 members of the control group had colds. What can you say about the DLS of these data, assuming that the supplement has no effect on whether or not an individual gets a cold.

CHAPTER 14. *HYPOTHESIS TESTING*

If we carry out an experimental procedure and make an observation; if we assume that a certain fixed model holds; and if we choose some metric to measure how far the observation is from giving strongest confirmation of the model, then, as we have seen above, we can calculate a descriptive level of significance (*DLS*). This *DLS* gives us an indication of *how good our model is*. A large *DLS* tends to confirm the model, while a small *DLS* suggests that the model may not be correct.

Sometimes, in statistics, we find that we have to make a *decision* as to whether or not a certain model is correct on the basis of an observation, and that we may then have to take some action on the basis of that decision. This action may take the form (to give only a few examples) of rejecting a shipment of goods as defective, of prohibiting the sale of a certain drug as harmful, of choosing a certain direction for further research, or of asserting and publishing a certain conclusion in a scientific paper.

In the present chapter, we shall limit ourselves to situations where we make a *two-way* decision of *either keeping or discarding a proposed model.* Decisions with more than two options can also arise in statistics; we treat them further in Chapter 20. (For example, a *three-way* decision might be: *keep model/ get more data/discard model*.)

The most common method for making a two-way decision about a model is to establish, ahead of time, a certain *cut-off* value α for the *DLS*. This value is called the *critical level of significance*. Then if the observed *DLS* falls above α we keep the model, while if if falls at or below α, we discard the model as incorrect. (If an observation has *DLS* $\leq \alpha$, we say that the observation is *statistically significant*.) The method is obvious, simple, and useful. How do we choose a value for α? This will depend upon the actions that we may take as a result of our decision, and upon the later consequences of those actions (where the decision proves to be good or bad). We discuss this further below. The values most commonly used in practice are $\alpha = 0.05$ and $\alpha = 0.01$.

Example. A certain experiment can result in success or failure We call the probability of success p. If we conduct a sequence of independent repetitions of this experiment, we have a sequence of Bernoulli trials. We take as our assumed model $p = \frac{1}{3}$. We take our critical level of significance to be $\alpha = 0.05$. We then carry out 100 trials and observe 45 successes. Does this lead us to keep the model or to discard it as incorrect? To answer this question, we calculate the *DLS* using the standard binomial metric $|X-np|$. Normal approximation gives us:

$$\underline{DLS} = 1-2A(z), \text{ where}$$

$$z = \frac{|X-np|-\frac{1}{2}}{\sqrt{np(1-p)}} = \frac{45-33.3-0.5}{\frac{10}{3}\sqrt{2}} = 2.38$$

Hence *DLS* = 1-2(0.4913) = 0.017. As this falls below α, we decide to discard the model $p = \frac{1}{3}$. Note that if we had taken $\alpha = 0.01$, we would decide to keep the model.

Some terminology. In a decision making procedure like the above, the model which we assume at the start is called the *hypothesis* or the *null hypothesis*. Let λ be the observed *DLS*. If $\lambda > \alpha$, we say that we *accept* or *continue to accept* the hypothesis, while if $\lambda \leq \alpha$, we say that we *reject* the hypothesis. The entire procedure of choosing α, obtaining an observation, choosing a metric, calculating λ, and making a decision is called a *test* or *hypothesis test*. Hypothesis testing is one of the most important and widely used methods in mathematical statistics.

Example (continued). When we have a hypothesis to be tested and a test procedure, then certain observations will lead us to accept the hypothesis and certain other observations will lead us to reject the hypothesis. The set of observations which lead to *rejection* of the hypothesis is called the *critical region* or *rejection region* of the test. The set of observations which lead to *acceptance* of the hypothesis is called the *acceptance region* of the test. In the case of the example above, with $p = \frac{1}{3}$ as null hypothesis, with $\alpha = 0.05$, with $|X-np|$ as metric, and with $n = 100$ trials for our observation, we get the critical region by noting that

$$1\text{-}2A(1.96) = 0.05 \quad \text{and setting}$$

$$1.96 = \frac{|X-np|-\frac{1}{2}}{\sqrt{npq}} \quad .$$

Solving for X, we get $X = 43.07$ and $X = 23.55$. Therefore the *critical region* consists of values of X such that

$$X \leq 23 \quad \text{or} \quad X \geq 44,$$

and the acceptance region is

$$24 \leq X \leq 43.$$

Remark. Note that even after we have fixed the hypothesis to be tested, we have a great deal of freedom in the way in which we design a test. In particular, the following aspects of the test remain to be determined: (i) the kind of observation to be made and the experimental procedure to be used in making it (for example, size and randomization of sample); (ii) the value of the critical level α; (iii) the metric to be used to get the DLS; (iv) possible use of approximations in calculating the DLS. In addition, it may be appropriate to use a composite model in a statistical test. Our decision will then be a decision as to whether the true model falls in the composite model or not. For such a test, we usually use a metric that is well-defined (gives a common DLS value) over all models in the composite model. The composite model itself is then called the null hypothesis for our test. Much ingenuity, art, and theory can go into determining these various aspects of a hypothesis test.

As an illustration of the various choices that can occur with (iii), consider the case of a binomial experiment and the hypothesis $p = \frac{1}{3}$. An observation is obtained for five successive runs of 100 trials each. We get 30, 44, 29, 38, 42 as the numbers of successes in the five runs. The theoretically expected result (TER) for a single run is 33.3. Various different metrics are possible. We could use $|A - 33.3|$ where $A = 36.6$ is the average of the five observed numbers, or we could use $|M - 33.3|$ where $M = 38$ is the median (middle) value of the observed numbers. Which choice is better? It is possible to show (for reasons to be given below) that the first metric is better in this case. It is also easier to calculate the DLS with the first metric; ease of calculation can be an important part of designing a test.

Examples. Contingency tables can be used in hypothesis tests. For example, consider two attributes A and B under a multinomial procedure (type α), and take as our hypothesis the composite model that the attributes are not associated. Fix $\alpha = 0.05$. If we obtain the table

29	24
5	10

then we get DLS = 0.14. (This example was calculated in Chapter 13.). Hence we continue to accept the hypothesis that the attributes are not associated.

On the other hand, the table for smoking/cancer was

50	950
10	8990

Here we saw DLS < 0.001, so that we would reject the hypothesis of non-association when $\alpha = 0.05$ (and also when $\alpha = 0.01$ or $\alpha = 0.001$).

Goodness-of-fit can also be the subject of hypothesis tests. In the example on page 333 on goodness-of-fit of a given set of observations to the standard normal curve, we would reject the standard normal curve as model in a test at critical level 0.05, but would continue to accept in a test at level 0.01. Similarly, in the example on page 337 on the goodness-of-fit of a given set of observations to some (unspecified) Poisson distribution, we would reject the composite model of all Poisson distributions in a test at level 0.05 but would continue to accept at level 0.01.

Meaning of the critical level α. Note that a hypothesis test does not treat the options of accepting and of rejecting in a symmetrical way. In effect, the test is biased in favor of the hypothesis, for we only reject when we get strong and convincing evidence that the model is wrong. When the evidence is uncertain, we continue to accept.

If we carry out a hypothesis test, we can reach a wrong conclusion in one of two different ways:

I. We can reject the null hypothesis when the null hypothesis is in fact true.

II. We can accept the null hypothesis when the null hypothesis is in fact false.

The critical level of significance α gives direct information about mistakes of type I. It tells us that if we repeat our entire test many times, and if the null hypothesis is in fact true, then we will make the type I mistake of rejecting our hypothesis about α of the time. α says nothing directly about how often mistakes of type II will occur when the null hypothesis is false. Thus when $\alpha = 0.05$, if the hypothesis is actually true, we can expect to make the mistake (of rejecting it) about $\frac{1}{20}$ of the time if we keep repeating the entire test procedure. The smaller the critical level α, the more cautious we are being about rejecting the null hypothesis if it is true, and the more convincing is the evidence that we require for rejection.

The power of a test. What about a mistake of type II? What is the chance of *accepting* the null hypothesis when it is in fact *false*? We first restate the question as a question about the opposite event.

What is the chance of *correctly rejecting* the null hypothesis when it is in fact *false*? This question cannot be answered *unless we pick some other model* $\underline{\mu_1}$ which we take as the true model and which we then use to calculate ρ = *the probability that we will correctly reject the original hypothesis when our test is carried out*. This probability ρ is called the *power* of the test *with respect to the alternative model* $\underline{\mu_1}$. If we assume that μ_1 is true, then $\beta = 1-\rho$ is the probability of making a mistake of type II.

Example. Assume again that we have Bernoulli trials with the null hypothesis that $p = \frac{1}{3}$. Our test, as before, is to observe 100 trials and use critical level $\alpha = 0.05$. Now assume that the null hypothesis is false and that the true model μ_1 is $p = \frac{1}{2}$. What is the power of our test with respect to the model $p = \frac{1}{2}$? We need to calculate the probability (under $p = \frac{1}{2}$) of rejecting $p = \frac{1}{3}$, that is to say of obtaining an observation that falls in the critical region of the given test for $p = \frac{1}{3}$. Recall from the example at the beginning of this section that we reject $p = \frac{1}{3}$ if the observed X is ≥ 44 or ≤ 23. We thus get our desired power by finding the probability of the event $(X \leq 23$ or $X \geq 44)$ for the alternate model $p = \frac{1}{2}$. Using normal approximation, we get this probability to be 0.903. Thus $\rho = 0.903$ for μ_1.

The power function. Let μ_0 be the model chosen as null hypothesis for some given test. Let μ be any possible model (either μ_0 or some alternative model to μ_0). Define $\rho(\mu)$ to be the probability of rejecting μ_0 given that μ is true. Then for any alternative model μ_1, $\rho(\mu_1)$ must be the power of the test with respect to the alternative model μ_1 (we called this value ρ above), while for μ_0, $\rho(\mu_0)$ must be ≈ the critical level α of the test. (It may be less than α if the observation can take on only separated discrete values. In the example above, $\rho(\mu_0)$ = the probability of the event $(X \leq 23$ or $X \geq 44)$ for the model $p = \frac{1}{3}$. Using normal approximation, we get $\rho(\mu_0) = 0.034 < \alpha = 0.05$.) The function ρ, defined in this way for all possible models μ (in the universe of models), is called the power function of the given hypothesis test. In the case of the parametric example above, the possible models correspond to possible values of p and can be represented as points in the interval from 0 to 1. Hence in this case the power function can be graphed. Its graph is given by the solid curve in Figure 14.1.

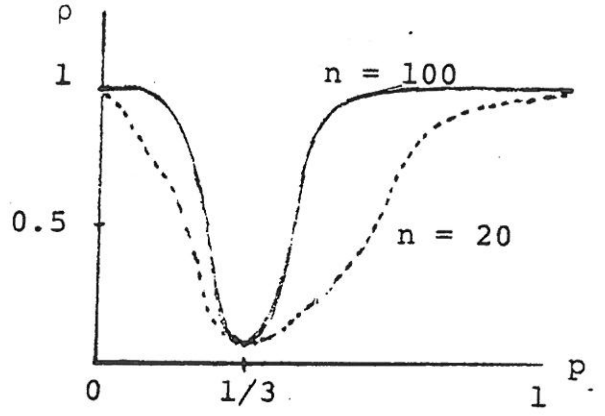

Figure 14.1

The power function can help decide what procedure to use for testing a given null hypothesis. For two tests with the same critical level, we will be inclined to use the test with the greater power function. (In the graph above, we have also drawn the power function for a test based on $n = 20$ trials and critical level $\alpha = 0.05$. As we would expect, the test with $n = 100$ is more powerful for all alternative models). We cannot always compare two tests in this way. It may happen that one test is more powerful for certain of the alternative models but that the other test is more powerful for the other alternative models. If a test with the same critical level α is more powerful for all alternative models, we say simply that it is a more powerful test. In our earlier illustration above of different metrics for five successive binomial runs of 100 trials each, the metric based on the average figure is superior to the metric based on the median figure because it yields a more powerful test.

Choosing the critical level of significance. While the critical level α gives the probability of a mistake of type I and does not say anything about the probability of a mistake of type II, changing the value of α will also affect the probability of type II mistakes. For example, if we raise the value of α for a given test, this can raise the entire power function. The best choice of α in a particular situation will depend upon the relative size of the penalties that we will suffer if we make a type I or a type II mistake. If the type I mistake is especially serious, we will be cautious and make α small. If a type II mistake is serious, this may lead us to a somewhat larger α. For example, if the null hypothesis is that a certain food additive is harmful, we will take α very small so that we will treat the additive as harmless (reject the hypothesis) only if we get very convincing evidence. On the other hand, if our null hypothesis is that an additive is harmless, we may wish to use a somewhat larger value of α; this means that weaker evidence can lead us to treat the additive as harmful (reject the hypothesis). Other features of test procedures are closely related to the choice of α. In particular, the advantages of a larger sample for an observation must be weighed against the cost in time and money of using a larger sample, and these must be balanced against the penalties for making mistakes of type I and type II. In mathematical statistics as developed prior to 1950 (classical mathematical statistics), considerations of cost and penalty were not usually made part of the mathematical formulation of a statistical problem. They were left as part of the art of a statistician rather than as part of the science. Since 1950 the approach known as statistical decision theory has attempted to make these considerations of cost and penalty a basic part of the mathematical formulation of statistical problems. Today, most statisticians accept and use many of the methods of decision theory, although certain aspects remain controversial. We shall consider statistical decision theory in Chapter 20. Until Chapter 20, the approach of our text will be primarily that of classical mathematical statistics.

As noted before, the most commonly used values of α are 0.05 and 0.01. These values are arbitrary, but they are important because they are so widely used. If we learn that someone has rejected a certain hypothesis on the basis of observed data, we can be almost certain that a value of $\alpha \leq 0.05$ has been used. We shall return to the choice of α in Chapter 20.

Designing tests: two-sided tests vs. one-sided tests. We begin with an example. Assume that at a certain hospital the rate of post-operative surgical infection is known to be $\frac{2}{3}$. That is to say, the probability of a patient remaining free from infection is $p = \frac{1}{3}$. Let us next assume that a new and more intensive procedure for sterilizing surgical bandages is introduced. (This new procedure, for example, might take the form of repeating the old sterilization procedure several times.) We now wish to test whether or not the new sterilization procedure for bandages gives a significant improvement in the infection rate. We take as our null hypothesis that $p = \frac{1}{3}$ (no improvement occurs), and we choose $\alpha = 0.05$. We observe 100 patients with the new bandages, and we find that X of them remain free from infection. How do we decide, from X, whether or not to reject the null hypothesis?

We begin by deciding on a <u>universe</u> of models for our test. We have an additional piece of information which occurs as a special feature of the problem: we know from the heory of infection that the new bandages will not <u>decrease</u> the value of p; thus the only possible alternative models to the null hypothesis $p = \frac{1}{3}$ are models with $p > \frac{1}{3}$. This is our universe of models. Our situation is said to be <u>one-sided</u> because the alternative models all lie on one side of the null hypothesis in our parametric universe. The situation in the earlier example, where alternative models could have either $p < \frac{1}{3}$ or $p > \frac{1}{3}$, is called <u>two-sided</u>.

What if we now observe $X = 15$? In the two-sided situation as we saw above, any observation ≤ 23 will lead us to reject the null hypothesis. Here, however, in the one-sided situation, we cannot reject the null hypothesis, for even though the observation $X = 15$ is highly unlikely under the null hypothesis, it is <u>even</u> more unlikely under any alternative model in the universe. If and when we observe $X = 15$, we must interpret this observation merely as an unlikely random occurrence rather than as evidence against the null hypothesis.

(Indeed, if the unlikely observation $X = 15$ occurs, we can view it as confirming the null hypothesis because it is less unlikely under the null hypothesis than under any other model in the universe.)

How do we formulate an appropriate test procedure? Since only values of $X > 33.3$ can cast doubt on the hypothesis $p = \frac{1}{3}$, we seek a critical region lying entirely to the right of 33.3 having total probability ≤ 0.05. Normal approximation shows us that under the null hypothesis, $P(X \geq 42) \leq 0.05$, but $P(X \geq 41) > 0.05$. Hence we take $X \geq 42$ as our critical region, and we reject the null hypothesis if we get $X \geq 42$. If the null hypothesis is in fact true, this test will lead us into an error of type I with probability ≤ 0.05. We call this test a <u>one-sided test</u>. Recall from our earlier example that the corresponding <u>two-sided test</u> has $X \leq 23$ or $X \geq 44$ as its critical region. Thus, for example, the observation $X = 43$ leads us to reject the null hypothesis $p = \frac{1}{3}$ under the one-sided test but not under the two-sided test. This can be explained intuitively, in part, by saying that in the one-sided case we have more information to start with about the possible models that can occur, and that we hence need less additional information in order to reject the hypothesis.

Another and more precise way to look at the difference between one-sided and two-sided tests is simply to note that we are using a different metric to calculate the <u>DLS</u> in the two situations. In the two-sided case above, we use the metric

$$s(X) = |X\text{-}np|;$$

in the one-sided case, we use the metric

$$s(X) = \frac{1}{2}(|X\text{-}np| + (X\text{-}np)) = \begin{cases} X\text{-}np & \text{when} \quad X > np; \\ 0 & \text{when} \quad X \leq np. \end{cases}$$

This is one of several choices of metric that can be used in the present example. How to choose the best metric for a one-sided universe is not always obvious. We shall return to this matter in Chapter 20 when we consider the general question of finding natural metrics in statistical problems. Note that $s(X) = X$ could also be used as a metric for the one-sided test above, since we could intuitively argue that $X = 0$, though highly unlikely, is the <u>most</u> strongly confirming observation of all (when it occurs) because it is so much <u>more</u> unlikely under any model (in the universe) other than the null hypothesis. (Both one-sided metrics give the same <u>DLS</u> values for $X > np$ and hence produce the same critical region for $\alpha = 0.05$.)

One-sided and two-sided tests also arise in connection with 2 x 2 contingency tables. We might, for example, have <u>no</u> previous information about the rate of infection in our hospital, and we might observe 34 patients with new bandages and 34 patients with old bandages. This might give us the following table

	No infection	Infection
New bandages	29	5
Old bandages	24	10

We take our null hypothesis to be that there is no difference between new and old bandages, that is to say, that there is no association in the table. We take our critical level to be 0.10. In an example in Chapter 13, by a large-sample calculation, we saw that this table has DLS = 0.14. However, this DLS was based on the CS-metric. In our present example, this metric treats observations that associate more infection with old bandages and observations that associate more infection with new bandages on the same basis. Because, under the theory of infection, we know that more infection cannot be associated with new bandages, we use instead a one-sided metric (for the same reasons as given in the previous discussion above). We take

$$s(\Omega) = \begin{cases} \chi^2 & \text{if} \quad a > E_1 , \\ 0 & \text{if} \quad a \leq E_1 , \end{cases}$$

where a is the count observed in the upper left cell in observed table Ω, and χ^2 is the value of the CS-metric for Ω.

It is possible to show, from the definitionsin Chapter 13, that the one-sided DLS of a contingency table is exactly one half its two-sided DLS. Thus, for the table given above, we get DLS = $\frac{1}{2}(0.14)$ = 0.07, and, since our test has critical level 0.10, we reject the null hypothesis. We conclude that the new bandages are significantly better.

Comment. One-sided tests should be used, if at all, with special care. When we limit a parametric universe to some subset of parameter values, we may be ignoring a slight possibility that one of the other parameter values in reality occurs. An extreme observation on the nonsignificant side (of the one-sided test) may well, in practice, lead us actively to explore the possibility that one of the parameter values omitted for the one-sided test has in fact occurred. For example, in the case above of 100 trials of patients with new bandages, if we actually observe $x = 5$ (95 infected patients) we will not, in practice, view this as confirming $p = 1/3$, but rather as suggesting that some unknown and unexpected source of infection (which we had ruled out in our one-sided universe) has entered the picture. The distinction between two-sided and one-sided tests is thus a somewhat arbitrary and artificial one. Often, what is needed is an appropriate metric that gives DLS values somewhere between the values from a one-sided universe and the values from symmetrical treatment of a two-sided universe. We come back to this question in Chapter 20.

Footnote. In the case of the roulette wheel data given in Exercise 10-18, the distinction between the DLS value in part (a) and the DLS value in part (b) can be described as a distinction between a one-sided situation and a two-sided situation. The DLS value obtained in (b) (for goodness-of-fit to a given Poisson distribution) can be shown to be approximately the same as the DLS value that would be obtained using the metric $|CS(\Omega)-m_o|$ where $CS(\Omega)$ is the CS-metric from (a) and m_o is taken so that $C_{37}(m_o) = 1/2$. Situation (b) can be described as two-sided, because the metric measures how far $CS(\Omega)$ varies on either side of m_o. Situation (a) can be described as one-sided, because the CS-metric measures how far $CS(\Omega)$ lies above 0. As we would expect from the symmetry of the chi-square curves for large d (where d = degrees of freedom), it is the case, when the DLS in (a) is $\leq 1/2$, that the DLS in (a) is approximately one half the DLS in (b).

Designing tests: randomization. A famous example of hypothesis testing in statistics is due to R. A. Fisher. In this example (the lady tasting tea), a woman claims that she can tell from tasting a cup of mixed tea and milk which of two different methods of pouring the tea has been used (milk poured first/milk poured last). The null hypothesis is taken to be that she cannot in fact discriminate. The procedure is to give her eight cups, four poured one way and four the other. She is told that there are four of each kind, and is asked to identify which four have milk first and which four have milk last. It is easy to see that the proper test for $\alpha = 0.05$ is to require that she correctly identify all eight. (We assume that the woman always identifies four cups of each kind. There are $\binom{8}{4} = 70$ possible identifications that she can make. Under the null hypothesis, these are equally likely. The probability of getting all cups correct is therefore $\frac{1}{70}$. The probability of making one mistake (interchanging one cup of each kind) is $\frac{16}{70}$. Since $\frac{17}{70} > .05$ and $\frac{1}{70} \leq 0.5$, we must require that all cups be identified correctly in order to have $\alpha = 0.05$. More formally, if R is the number of cups identified correctly, we can view the test as a one-sided test and take

$$s(R) = \begin{cases} R - 4 & \text{if} \quad R \geq 4 \\ 0 & \text{if} \quad R < 4 \end{cases}$$

as our metric. The DLS for $R = 8$ is then $1/70 = 0.01$, while the DLS for $R = 6$ is $17/70 = 0.24$.

Several difficulties now present themselves. The first is that the woman's judgments may be affected by the order in which the examples of the two preparations are given to her. (It could conceivably be the case, for example, that the woman's taste is always affected by the first cup so that she thinks that the next three cups are identical with the first cup.) We can take care of this difficulty by using a random digit table to randomize the order in which the cups are given to her. This randomization becomes an integral part of the entire test procedure. If we repeat the test, we use a new randomization. This means that if we repeat the entire test many times, each order will occur equally often. If the null hypothesis is true, then, no matter how the woman is influenced by the order of the cups, she will, on the average, get all eight answers correct about $\frac{1}{70}$ of the time. Making this randomization part of our experiment serves to protect the calculated DLS values from any accidental ways in which the order of the samples might affect the woman's statements.

A second difficulty might arise if we did not have eight identical cups to pour the tea and milk into. With limited resources, for example, we could conceivably find ourselves forced to use a mixture of paper cups, china cups, and metal cups. Which cups should we use for which kind of tea? It might be that the woman's judgment is affected, either physically or psychologically, by the material of the cup she is drinking from. She might, for example, always tend to believe that tea in a metal cup had had the milk poured first. Can we take care of this difficulty, or must we wait until we can get eight identical cups? Randomization again resolves our difficulty. We use random digits to assign the two methods of pouring to the eight cups (four cups for each method), and we make this randomization an integral part of our entire test procedure. Now, if we repeat the test, we repeat the randomization anew. As before, this ensures that in many repetitions of the test, if the null hypothesis is true, the woman will get all eight identifications correct about $\frac{1}{70}$ of the time. Again, we have protected the calculated DLS values.

Is there then any advantage in using identical vessels? Yes, there is a clear advantage, but this advantage does not have to do with the critical level of the test. It has to do, rather, with the power of the test. The test with identical cups may well be more powerful because it is less confusing to the woman. That is to say, it may make it easier for the woman to demonstrate her ability if she actually has it (null hypothesis false). The DLS values on the other hand, which we have protected by randomization, have only to do with the case where the woman in fact has no ability (null hypothesis true).

Fallacies in hypothesis testing. Conceptual and mathematical errors can occur in hypothesis testing. Such errors sometimes arise when an investigator has strong subjective beliefs about the truth or falsity of the null hypothesis and allows these beliefs to affect the experimental and mathematical procedures followed.

Some errors are obvious as errors. These include: (a) outright discarding of data that is contrary to the investigator's beliefs; (b) use of observed data to help formulate a null hypothesis which is then tested by the use of the same data; (c) selective reporting of significant results from a larger body of results that includes many non-significant results. Questions about whether or not such obvious errors as these have been made can occasionally cause controversies over the interpretation of statistical data in particular cases.

There are other, more subtle errors that can be equally serious and that often go unrecognized. A typical form of such an error is as follows. A hypothesis is tested by making successively larger observations where all the data from one observation are included in the next larger observation. Thus no data are discarded. The investigator calculates the DLS for each successive observation. If and when the investigator finds a DLS below the chosen critical value α, the investigator concludes that the null hypothesis is false and rejects it.

This error can arise in a natural way, for example, when an investigator uses contingency tables to test for the possible effect of some drug on human subjects. In order to minimize the number of subjects used, the investigator starts with a small sample and then includes it in successively larger samples while seeking a contingency table with DLS $\leq \alpha$.

Such a procedure is, unfortunately, invalid. If, for example, an investigator makes five successively larger contingency tables, and tests each for significance at $\alpha = 0.05$, it is possible to show, (by methods more advanced than those covered in this book) that if the null hypothesis is really true, then the probability of getting at least one significant result out of the five can be as much as 0.14. More generally, it is possible to show that for any critical level α, no matter how small, it is virtually certain that if we carry out far enough a single open-ended procedure of going to larger and larger samples, we will eventually find a significant contingency table. This surprising result follows from theoretical consideration of the fluctuations that can occur as data are progressively accumulated from a probabilistic experiment.

There is one satisfactory way of avoiding this error. The investigator must (i) fix ahead of time the number of successive samples that will be used, and (ii) make a compensating adjustment in the size of the α used in the individual samples. For example, if the number of samples is five, the critical level for each sample must be 0.015 in order to ensure that the entire procedure of testing five samples have a critical level of 0.05. (Calculation of this compensating value again requires more advanced theory than is covered in this book.)

Errors of this kind are not uncommon in biological and medical studies. They have been alleged to occur in hypotheses testing related to extra-sensory perception.

Theoretical note. At the end of Chapter 13, we saw that in a type α or type β contingency table, $\underline{DLS}_t$ is always less than or equal to the conditional $\underline{DLS}_c$ calculated by the small sample method. Thus in a hypothesis test for which non-association is the null hypothesis, with $\alpha = 0.05$, we can reject when $\underline{DLS}_c \leq 0.05$, for we know that if $\underline{DLS}_c(\Omega) \leq 0.05$, then $\underline{DLS}_t(\Omega;\mu)$ must also be ≤ 0.05 for all models μ of non-association. In this case, we cannot speak of a unique probability for a type I error. We can only say that for each model in the composite model, the probability of a type I error, when that model is the correct model, is ≤ 0.05.

The small sample method was developed in Chapter 13 for type γ tables, and it gives us a hypothesis test for type γ tables. We have now seen that the inequality $\underline{DLS}_t \leq \underline{DLS}_c$ allows us to extend the application of this test to type α and type β tables. A test which remains valid when given wider applicability (as in this case) is said to be robust.

Example. In a double binomial experiment of the kind described in Chapter 11, a control group of 4 rats and an experimental group of 4 rats are exposed to infection. All four of the control group show signs of infection, but none of the experimental group does. Is this result statistically significant at critical level $\alpha = 0.05$? The observed table is

4	0
0	4

The small sample method gives $\underline{DLS}_c = \dfrac{2\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \dfrac{2}{70} = 0.03.$

Hence the observation is significant. If the table were

3	0
0	3

we would get $\underline{DLS}_c = \dfrac{2\binom{3}{3}\binom{3}{0}}{\binom{6}{3}} = \dfrac{2}{20} = 0.10,$ and this table would not be significant at critical level $\alpha = 0.05$. (On the other hand, in a one-sided test this table would be significant, since its $\underline{DLS}_c$ would be exactly 0.05.)

SOME EXERCISES ON CHAPTER 14

14-1. Fred has a die he believes may be loaded in favor of the side marked "six". He tosses it 4 times and gets 3 "sixes". Using the 5% level of significance, do these results cause you to reject the null hypothesis $p = \frac{1}{6}$?

14-2. Mr. Williams played 5 hands of bridge one evening and got no aces 4 times. He complained of poor shuffling. Assuming good shuffling, the probability p of getting at least 1 ace, on any 1 deal, is 0.7 (approximately). Are 4 no-ace hands out of 5 hands enough to reject the null hypothesis $p = 0.7$ at the 5% level of significance? (Use the binomial formula.)

14-3. A manufacturer of light bulbs says that only 10% of the frosted bulbs he manufactures have defective frosting, and that these defective bulbs occur at random during manufacture. A carton of 4 of his bulbs was purchased and 2 of these had defective frosting. Would you reject his claim at the 1% level of significance?

14-4. In a coffee-tasting experiment a subject tastes each of 10 pairs of cups of coffee and decides for each pair which cup contains the instant rather than the percolated coffee. The experimenter decides to call a person a "taster" if he or she decides correctly in at least 8 out of 10 pairs; otherwise the person is called a "nontaster". Regarding this operation as a test of significance: (a) What is the null hypothesis? (b) What are the alternative hypotheses? (c) What is the level of significance? (d) If a subject has probability 0.8 of correctly calling a pair, what is the chance the subject will be called a "taster"?

14-5. In a random sample of 900 cars, 200 are observed to be station wagons. Test the hypothesis that 15% of the underlying population are station wagons. Use critical level $\alpha = 0.01$.

14-6. Let X be the number of traffic accidents in a given town in a week. X is observed for 20 different weeks and the following table is obtained:

X:	0	1	2	3
Number of weeks with X accidents:	5	9	5	1

a. Fit a Poisson distribution to these observations.

b. Assuming that X *has* a Poisson distribution, use chi-square approximation to find the *DLS* of the observed data. (*Note*: Please pool observations with $X \geqslant 2$. Use only two significant figures in calculations.)

c. Use the observed data to test, at critical level $\alpha = 0.05$, the hypothesis that X has a Poisson distribution.