CHAPTER 11. CHI-SQUARE APPROXIMATION.

The chi-square curves. In Chapter 6, we saw how area under the standard normal curve can be used to get approximate probabilities for a binomial distribution. We were then able, in Chapter 10, to use the standard normal curve to get approximate values for the descriptive level of significance with a binomial model. We now see how certain other curves can be used, in a similar way, to get approximate values for the descriptive level of significance with a multinomial model and the CS metric.

These curves are called the chi-square curves. For each positive integer d, there is a corresponding standard curve called the chi-square curve with d degrees of freedom. This curve has the equation

$$\psi_d(u) = A_d u^{\left(\frac{d}{2} - 1\right)} e^{-\frac{u}{2}},$$

where the constant A_d is chosen so that the total area under the curve ψ_d is 1. (In fact, for d odd, $A_d = \dfrac{2^{\frac{d}{2}-1}\left(\frac{d}{2}-\frac{1}{2}\right)!}{\sqrt{\pi}\ (d-1)!}$; and for d even, $A_d = \dfrac{1}{2^{\frac{d}{2}}\left(\frac{d}{2}-1\right)!}$.) The curves are shown in the following figures for $d = 1,2,3,5,10,$ and 20.

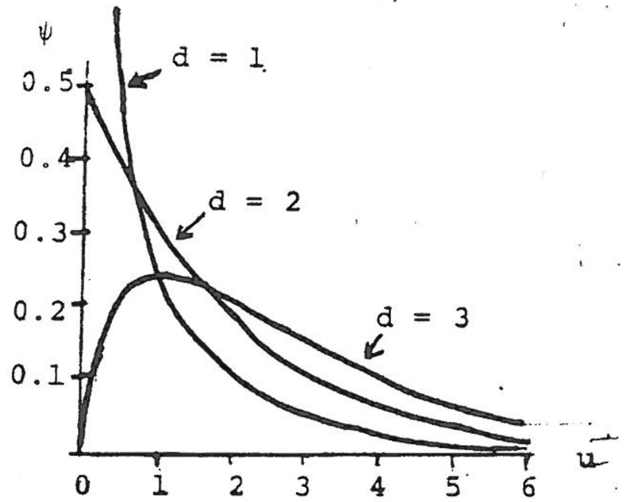

Figure 11.1

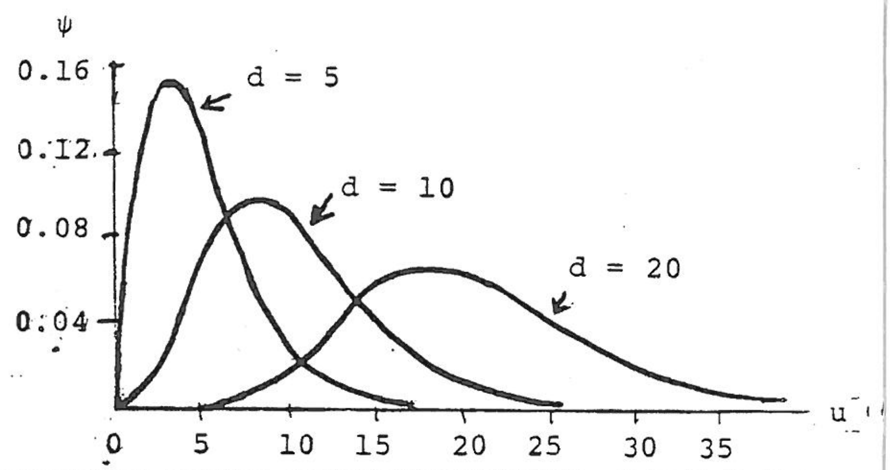

Figure 11.2

(Footnote. Several of these curves are familiar in physics. The curve ψ_2 is a curve for exponential decay (from an initial value of 0.5, where u measures time, and the half-life is $u = 2 \log 2$). The curve ψ_3 is Maxwell's distribution for the translational kinetic energy of particles in a gas at a given temperature in three-dimensional space. Here u measures kinetic energy in units of $\frac{1}{2}kT$ (for T = absolute temperature and k = Boltzmann's constant). More generally, ψ_d gives the Maxwellian distribution for kinetic energy of a system with d physical degrees of freedom, where u measures kinetic energy in units of $\frac{1}{2}kT$.)

Tables. Areas under the chi-square curves are given in tables. Usually, these tables give areas determined by various values of u for each $d \leq 30$. For $d > 30$, the chi-square curve ψ_d can be approximated by the standard normal curve in a way that we describe later in this chapter. The quantity usually given in tables is what we shall call $C_d(u_o)$. This is the area lying to the right of a vertical line at u_o, as shown in Figure 11.3

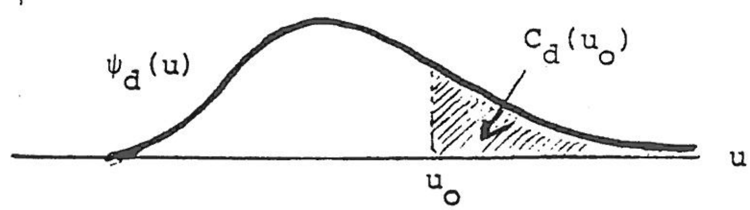

Figure 11.3

Tables for this area are sometimes given in inverse form (for reasons that we shall see in Chapter 14), where, instead of giving the value of C_d for certain chosen values of u, the table gives values of u for certain chosen values of the area C_d. For example, a table for ψ_2 might be given in the form:

$C_2(u)$	0.95	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05
u	0.10	0.21	0.45	0.71	1.02	1.39	1.83	2.41	3.22	4.61	5.99

If we wish to know $C_2(3)$, we can get an approximate value by locating the value 3 between 2.41 and 3.22 and then finding the value which falls at a corresponding place between 0.3 and 0.2. By this process of linear interpolation in the above table, we find $C_2(3) \approx 0.3 - 0.1 \dfrac{3 - 2.41}{3.22 - 2.41} = 0.23$. Tables of chi-square areas for $d \leq 30$ are given, in inverse form, at the end of this chapter.

Expectation and variance. Let X be a random variable whose distribution is given by $f(x) = \psi_d(x)$, the chi-square curve with d degrees of freedom. It is possible to show (see the Appendix to this chapter) that $E_X = d$ and $V_X = 2d$. Hence the standardized form of such a chi-square variable is given by $\dfrac{X-d}{\sqrt{2d}}$.

Chi-square approximation. We first state the basic fact without proof. Let a multinomial experiment with c categories be given, and let a fixed model for this experiment be assumed. If an observation is obtained, and if the value χ_o^2 of the CS metric is got for this observation, then the value of the descriptive level of significance $P(\chi^2 \geq \chi_o^2)$ is approximately given by the area $C_d(\chi_o^2)$, where $d = c-1$.

Example. Take the Boston-Montreal example in Chapter 10. Our model was a multinomial experiment of 3 categories, with $n = 10$ and fixed probabilities $p_1 = 0.5$, $p_2 = 0.3$, $p_3 = 0.2$. For the observed result $(5,1,4)$, we saw in Chapter 10 that $\chi_o^2 = 3.33$. To get an approximate value using chi-square curves, we use the curve with $3 - 1 = 2$ degrees of freedom, and we get

$$P(\chi^2 \geq 3.33) \approx C_2(3.33) = 0.19 .$$

As we saw in Chapter 10, the exact value for this descriptive level of significance is 0.21.

Accuracy. The chi-square curves give an accuracy to two decimal places if the expected number in each category (called $E_1, E_2, \ldots E_c$ in Chapter 10, where $E_1 = np_1$, $E_2 = np_2, \ldots$) is ≥ 10. If $E_1, E_2, \ldots$ are each ≥ 5, we get close to two-decimal accuracy when $c \geq 5$, and even if some values of $E_1, E_2, \ldots$ are smaller than 5, and $c < 5$, the curves give close to two-decimal accuracy for values of $u = \chi_o^2$ which come at the right hand end of the curve (that is, which give small descriptive levels.) In the example of Chapter 10, where $E_1 = 5$, $E_2 = 3$, $E_3 = 2$, the observation (3,2,5) gives $\chi_o^2 = 5.63$; we thus have, for $P(\chi^2 \geq \chi_o^2)$, the exact value 0.06, while the chi-square curve with 2 degrees of freedom also gives the value 0.06. With the observation (4,4,2), however, where $\chi_o^2 = 0.53$, the exact value for $P(\chi^2 \geq \chi_o^2)$ is 0.92, while the chi-square curve gives the value 0.77.

Example. A die is rolled 100 times, and the number of occurrences of each of the six different individual outcomes is recorded. We get the following observation: 12 ones, 15 twos, 28 threes, 15 fours, 18 fives, 12 sixes. Assume the equiprobable model. What is the descriptive level of significance of this observation?

We have a multinomial experiment with 6 categories and $n = 100$. For our model, we have $p_1 = p_2 = \ldots = p_6 = \frac{1}{6}$. The theoretically expected result is $E_1 = E_2 = \ldots = E_6 = 100 \cdot \frac{1}{6} = 16.7$. We therefore get, as the value of our CS metric,

$$\chi_o^2 = \frac{(12 - 16.7)^2}{16.7} + \frac{(15 - 16.7)^2}{16.7} + \frac{(28 - 16.7)^2}{16.7} + \frac{(15 - 16.7)^2}{16.7} + \frac{(18 - 16.7)^2}{16.7}$$

$$+ \frac{(12 - 16.7)^2}{16.7} = 1.32 + 0.17 + 7.65 + 0.17 + 0.10 + 1.32 = 10.73.$$

(This is a case where fractional values appear in the theoretically expected result even though they cannot occur in an actual observation.) Using a table for the chi-square curve with $6 - 1 = 5$ degrees of freedom, we get the descriptive level of significance for this observation: $P(\chi^2 \geq 10.73) \approx C_5(10.73) = 0.06$. Thus, if the assumed model is correct, we would expect an observation as extreme as this only about 6% of the time. Note that in this example the expected values were ≥ 10 (they were all = 16.7), hence we can be sure of two-decimal-place accuracy in our result.

Finite models. If we have a basic experiment and take a model whose sample space is finite, and if we do n independent trials of the basic experiment and observe the number of times that each sample point occurs in these trials, then our over-all procedure is a multinomial experiment. We can calculate the DLS (under the CS metric) for our observation and assumed model by the above method. If the DLS is large, we may say that the observation "agrees with" or "fits" the assumed model. If the DLS is small, we may say that the observation "casts doubt" on the assumed model. If the DLS is very small, it may lead us to reject our assumed model as a useful picture of the experiment. (We look at this matter further in Chapter 14.)

Pooling. In using chi-square approximation to calculate the DLS, we may find that for some sample point x (in the underlying sample space of a single trial), the expected number of occurrences $E_x = np_x$ is small. Since a small value for E_x will reduce the accuracy of chi-square approximation, we may wish to take those points for which E_x is small and group them, for the purposes of our over-all multinomial model, with other points as a single category (or perhaps as several categories) for which the expected number of occurrences is not small. (Usually, if we wish to be certain of two-decimal place accuracy, we take "small" to mean less than 4 or 5.)

For example, if we have 50 trials and a model with six sample points (for each trial) as given in the first two columns of the following table, then the values E_x are as given in the third column. In order to get more accurate chi-square approximation, we group the first and second points together as a single category and the fifth and sixth points together as a single category. We now have a new form of multinomial experiment and model with 4 categories and can proceed with chi-square approximation as before. We speak of the new form of the experiment and model as pooled, because we have grouped together several of the categories in the original experiment and model.

x	p_x	E_x	O_x	$\hat{E}$	$\hat{O}$
1	0.02	1	2	10	6
2	0.18	9	4		
3	0.30	15	21	15	21
4	0.30	15	14	15	14
5	0.19	9.5	0	10	9
6	0.01	0.5	9		

The column under $\hat{E}$ gives the expected number of occurrences for each category in the pooled experiment. Let the column under O_x give the original observation. Then the column under $\hat{O}$ gives the corresponding observation for the pooled multinomial experiment. We obtain

$$\chi_o^2 = \frac{(6-10)^2}{10} + \frac{(21-15)^2}{15} + \frac{(14-15)^2}{15} + \frac{(9-10)^2}{10}$$

$$= 4.17;$$

and $\underline{DLS} = P(\chi^2 \geq \chi_o^2) \approx C_3(4.17) = 0.25.$

Chi-square approximation thus gives us the DLS value for the new pooled multinomial model. This value may be less useful than the DLS value for the original multinomial model, since information is lost when we pool. For example, in the Boston-Montreal example with the observation (5,1,4), if we pooled the losses and ties we would get

<table>
<tr><th>x</th><th>p_x</th><th>E_x</th><th>O_x</th><th>$\hat{E}$</th><th>$\hat{O}$</th></tr>
<tr><td>win</td><td>0.5</td><td>5</td><td>5</td><td>5</td><td>5</td></tr>
<tr><td>lose</td><td>0.3</td><td>3</td><td>1</td><td rowspan="2">5</td><td rowspan="2">5</td></tr>
<tr><td>tie</td><td>0.2</td><td>2</td><td>4</td></tr>
</table>

with $\chi_o^2 = 0$ and DLS $= P(\chi^2 \geq 0) = 1$. This DLS for the pooled experiment tells us much less than the value 0.19 obtained by chi-square approximation for the unpooled experiment, even though the approximated DLS did not have two-decimal-place accuracy. (Recall that the exact value of the DLS in this unpooled case was 0.21.) It is possible to show, when pooling is used in this way (in circumstances in which chi-square approximation can be applied), that the DLS (under the CS metric) after pooling is always bigger than, or approximately as big as, the DLS of the same observation before pooling. This tells us that <u>if an observation has a small DLS after pooling, its DLS for the original unpooled experiment must also be small</u>. Thus pooling is useful because it helps us to identify certain observations (but not necessarily all observations) for which the unpooled DLS is small. (If we actually make one of these observations, it will therefore cast doubt on the assumed model.)

It should be emphasized that the pooling of categories is carried out solely on the basis of the expected numbers of occurrences under the chosen model. The observed numbers of occurrences should be disregarded in pooling. Otherwise, as we shall see in Chapter 14, the DLS may fail to have the desired meaning.

Infinite discrete models. If we have a basic experiment and take a model whose sample space is infinite and whose probability function assigns a positive probability value to each sample point (a Poisson distribution would be such a model), then, as we saw in Chapter 2, we say that we have an infinite discrete model. If we do n independent trials of the basic experiment and observe the number of times that each sample point occurs in those trials, then we can use a generalized CS metric, in much the same way as we used the CS metric for a multinomial experiment, by forming an infinite sum of the $\frac{(O-E)^2}{E}$ terms for the individual sample points. We shall also refer to this generalized metric as a CS metric.

Example. We know that the number of almonds in a certain brand of chocolate bar averages 5, and we assume, as a model, that the number of almonds per bar follows a Poisson distribution with $m = 5$. We examine 40 bars and get the following multiple observation:

x = number of almonds	0	1	2	3	4	5	6	7	8	9	10	...
O_x = observed number with x almonds	0	1	5	11	15	3	3	2	0	0	0	...

We can evaluate the CS metric for this model and this multiple observation by constructing the following table, where the second column is obtained from the Poisson table with $m = 5$, the third column gives the theoretically expected result, the fourth column gives the observation, and the last column gives the corresponding term for the CS metric.

x	p_x	$E_x=np_x$	O_x	$\frac{(O_x-E_x)^2}{E_x}$
0	0.007	0.3	0	0.3
1	0.034	1.4	1	0.1
2	0.084	3.4	5	0.8
3	0.140	5.6	11	5.2
4	0.175	7	15	9.1
5	0.175	7	3	2.3
6	0.146	5.8	3	1.4
7	0.104	4.2	2	1.2
8	0.065	2.6	0	2.6
9	0.036	1.4	0	1.4
10	0.018	0.7	0	0.7
⋮	sum = 0.014	sum = 0.6	⋮	sum = 0.6

Thus the value of the metric for this observation is $\chi_o^2 = 0.3 + 0.1 + 0.8 + 5.2 + \ldots = 25.7$. (Note that the term in the last column for each sample point x for which $O_x = 0$ is just the value of E_x for that point.)

Given an experiment and model, and given a multiple observation, this metric determines a DLS value (according to the basic definition of DLS). This DLS value can be approximated by pooling sample points in order to form an appropriate finite space and then applying chi-square approximation to the multinomial experiment and model that result from this pooling. (We do not prove this here.)

Example (continued). We pool the categories at the two ends of the distribution, by combining them to form single categories for which the expected value is ≥ 5. This gives the following table with six categories:

x	$\hat{E}$	$\hat{O}$	$\frac{(\hat{O} - \hat{E})^2}{\hat{E}}$
≤ 2	5.1	6	0.2
3	5.6	11	5.2
4	7	15	9.1
5	7	3	2.3
6	5.8	3	1.4
≥ 7	9.5	2	5.9

The last column of this table gives the individual terms for χ_o^2. Summing the last column, we get $\chi_o^2 = 24.1$. Using the chi-square curve with $6 - 1 = 5$ degrees of freedom, we get our descriptive level of significance: $P(\chi^2 \geq 24.1) \leq 0.001$. Thus our multiple observation suggests very strongly that our original model is not a good one.

The fact that the model above was a Poisson distribution does not have special importance. Any infinite discrete distribution could have been our model, and we could have found the DLS for a multiple observation under that model in the same way, using pooling, the CS metric, and chi-square approximation. A DLS found in this way is said, in statistics, to measure the goodness-of-fit of the multiple observation to the chosen model.

Continuous models. If we have a basic experiment and a continuous probability space as model, and if we perform n independent trials of the experiment, then we can calculate a DLS in a similar way. We take the sample space of the model (usually an interval of real numbers) and divide it into a finite number of subsets (usually a finite number of subintervals). We treat each subset as a category and then view n repeated independent trials of our basic experiment as a multinomial experiment of n trials with those categories. (We are, in effect, pooling all the points in each subset of the continuous probability space to form a point in a new finite probability space.) We then get the DLS, for the given multiple observation, in the usual way. This DLS is said to measure the goodness-of-fit of the multiple observation to the chosen continuous probability space.

Example. An experiment yields as its outcome a real number. In 20 trials, we observe the following values to one decimal place accuracy: -1.0, 1.2, 1.9, 0.4, 1.0, -2.0, -2.1, 1.7, -3.0, 1.4, -1.0, 0.4, 2.5, -2.8, -0.5, 1.5, 1.8, 3.7, 0.2, 1.7. Assume, as model, a continuous probability space with the standard normal curve as its probability density function. To find the goodness-of-fit of the observation to the chosen model (in terms of a DLS value), we can proceed as follows. As before, in the case of pooling for a discrete probability space, we choose the subsets of the given continuous probability space so that the expected number of observations in each subset will be ≥ 4 or 5. We therefore take the four intervals $(-\infty,-0.68)$, $(-0.68,0)$, $(0,0.68)$, and $(0.68,\infty)$. From the table of normal curve areas, we know that each of these intervals has, under the assumed model, probability of about 1/4 of occurring on a single trial. Hence we can view the entire experiment of 20 trials as a multinomial experiment of four categories with $n = 20$ and $p_1 = p_2 = p_3 = p_4 = 1/4$. We can now make the following table.

Subset	$\hat{E}$	$\hat{O}$	$\left(\frac{\hat{O}-\hat{E}}{\hat{E}}\right)^2$
$(-\infty,-0.68)$	5	6	1/5
$(-0.68,0)$	5	1	16/5
$(0,0.68)$	5	3	4/5
$(0.68,\infty)$	5	10	25/5

Summing the last column, we get $\chi_o^2 = 9.2$. Using the chi-square curve with $4 - 1 = 3$ degrees of freedom, we get our <u>DLS</u>:

$$P(\chi^2 \geq 9.2) \approx 0.03.$$

The set of all models of a given kind. In the previous example with the chocolate almond bars, we were given the specific Poisson model with $m = 5$. We then observed 40 bars, and we found the DLS of this observation on the basis of this given model. What if we were not given the specific value $m = 5$, but were instead merely asked to find how well the observation confirms that some Poisson model (from the set of all Poisson models) holds? To get a DLS, we need a model μ and a metric s. Let us begin with the metric s. As we shall now see, the metric can be described without reference to a particular chosen model. Our metric s is simply the following. We take the observed data and fit a Poisson distribution to it in the usual way by letting m (the parameter in the Poisson distribution) be the average of the observed values. We then calculate the familiar CS metric for the observed data from this fitted model. We now define the value of the metric s to be the value obtained for this CS metric.

Example. In the case of the chocolate bars, we observed an average number of almonds per bar = 3.8. Hence the fitted Poisson distribution has $m = 3.8$. Thus the value of the CS metric for these data from this fitted model would be obtained from the following table by summing the last column. This is then the value that we take as the value for our new metric s. (The p_x column in this table is obtained from the Poisson formula $\frac{e^{-m}m^x}{x!}$ with $m = 3.8$.)

x	p_x	E_x	O_x	$\frac{(O_x - E_x)^2}{E_x}$
0	0.022	0.9	0	0.9
1	0.085	3.4	1	1.7
2	0.162	6.5	5	0.4
3	0.205	8.2	11	1.0
4	0.194	7.8	15	6.7
5	0.148	5.9	3	1.4
6	0.094	3.8	3	0.2
7	0.051	2.0	2	0.0
8	sum = 0.039	sum = 1.6	0	sum = 1.6
9			.	
10			.	
.			.	
.			.	
.			.	

We thus obtain $s = 0.9 + 1.7 + \ldots = 13.9$.

Now, if we choose a model μ, we can, in the usual way, define a DLS (under this model) for the metric s. (μ can be any model whatever for an experiment with the sample space $\{0,1,2,\ldots\}$.) We would, of course, expect the DLS value we get to depend upon which model we use. It is a surprising and profound theoretical fact (which we do not prove here) that (i) under any two Poisson models (and for a sufficiently large number n of independent trials) the DLS values (under the metric s) for the same observation are approximately the same. Thus if we use the model $m = 10$ or the model $m = 5$ or the model $m = 3.8$ with the above observed data, the above metric s always gives us approximately the same DLS value (which turns out to be about 0.02.)

It is also a theoretical fact that (ii) an approximate value for this DLS can be found by following the usual techniques for chi-square approximation (with pooling of sample points in the fitted distribution as may be necessary), except that the number of degrees of freedom must be smaller by 1 than before. (Footnote. In (i), "sufficiently large n" can be taken to mean "when n is large enough so that the usual conditions for accuracy of a chi-square approximation apply.")

Example (continued). As before, we take those points x for which E_x is smaller than 4 or 5 and pool them with other points. We get:

x	$\hat{E}$	$\hat{O}$	$\frac{(\hat{O}-\hat{E})^2}{\hat{E}}$
≤ 1	4.3	1	2.5
2	6.5	5	0.4
3	8.2	11	1.0
4	7.8	15	6.7
5	5.9	3	1.4
≥ 6	7.4	5	0.7

The last column of this table gives the individual terms for a value of χ_o^2 for this new finite experiment and model. Summing the final column, we get $\chi_o^2 = 12.7$. The finite multinomial experiment represented by the table above has six categories. We would normally use a chi-square curve with $6 - 1 = 5$ degrees of freedom for approximating the DLS of an observation for such an experiment. Fact (ii) above, however, tells us that we must subtract 1 further degree of freedom. Hence we use the curve with $6 - 1 - 1 = 4$ degrees of freedom. From the table for this curve, using $\chi_o^2 = 12.7$, we obtain the DLS value 0.02. The DLS value obtained in this way has a meaning similar to the meaning that we have previously given DLS values. This DLS holds (and is small) for all Poisson models. Hence it strongly suggests that the observation does not confirm the possibility that any Poisson model holds.

The above procedure, for calculating the common DLS of an observation for the set of all Poisson models (under the metric s), is simple and easy to use. Analogous procedures can also be used with certain other sets of models and certain other metrics (analogous to s) for which analogous versions of facts (i) and (ii) above hold. We shall sometimes refer to such a procedure as using the observation to indicate a model in a CS-metric-DLS calculation. We shall give further consideration to the concepts and theory behind such procedures in the Appendix to this chapter. Some of these concepts were suggested in comment (2) at the end of Chapter 10. We also comment further in the Appendix on the subtraction of further degrees of freedom in the chi-square approximation. This reduction corresponds to the intuitive fact that in allowing the observation to help indicate the model that it will be tested against, we have in a sense favored the observation; hence, in calculating a DLS, we must make up for this favoring by judging more severely. The reduction in degrees of freedom gives smaller DLS values and hence makes a severer judgment.

The reader should note that the metric s above has a different purpose than the particular metrics introduced in Chapter 10. The metrics in Chapter 10 were designed to measure confirmation of a single given model. The metric s above is designed to measure confirmation of a certain set of models (the set of all Poisson distributions). It is a metric of the kind described in comment (2) at the end of Chapter 10.

Example. Here is a second example of a DLS calculation of the above kind. We examine 40 chocolate almond bars of a different brand and get the data:

x = number of almonds	0	1	2	3	4	5	6 ...
O_x = observed number with x almonds	6	9	15	6	2	2	0 ...

How well does this observation confirm that some Poisson model holds? Here the observed average of almonds per bar is 1.88, and, taking a Poisson distribution with $m = 1.88$ as our indicated model, we get the following table:

x	p	$\hat{E}$	$\hat{O}$	$\frac{(\hat{O}-\hat{E})^2}{\hat{E}}$
0	0.153	6.1	6	0.0
1	0.288	11.5	9	0.5
2	0.270	10.8	15	1.6
3	0.168	6.7	6	0.1
≥ 4	0.121	4.9	4	0.2

Summing the final column, we get $\chi_o^2 = 2.4$. Using the chi-square curve with $5 - 1 - 1 = 3$ degrees of freedom, we get DLS = 0.49. Thus our observation of the second brand of bar suggests that it may well follow a Poisson distribution.

Normal approximation. For $d > 30$, the standard normal curve can be used to give areas under the chi-square curves as follows

$$C_d(u) \approx \text{Normal Area}_z^\infty$$

$$\text{where } z = \frac{u - d}{\sqrt{2d}} .$$

Thus $C_d(u) = \frac{1}{2} - A(\frac{u-d}{\sqrt{2d}})$ for $u > d$, $C_d(u) = \frac{1}{2} + A(\frac{d-u}{\sqrt{2d}})$ for $u < d$.

(This approximation gives 2 decimal-place accuracy except for $30 < d \le 350$ and values of u in the interval $d \pm \sqrt{d}$, where the approximation may only be good to within 0.03.)

Assume, for example, that we have obtained data from an experiment which we believe to be Poisson. Assume further that we have fitted a Poisson distribution to the observed data by using the observed average. Finally, assume that we have $c = 40$ categories and that we get $\chi_o^2 = 50.3$. What is the descriptive level of significance? In this case we need $d = c - 2 = 40 - 2 = 38$. Thus we need Normal Area_z^{∞} for $z = \frac{50.3 - 38}{\sqrt{76}} = 1.41$. Then, from tables, we have

$$\underline{\text{DLS}} = P(\chi^2 \ge \chi_o^2) = \frac{1}{2} - A(z) = \frac{1}{2} - 0.42 = 0.08.$$

Remark. It is useful to memorize several simple facts about the chi-square curves and their tables: (1) For the chi-square curve with d degrees of freedom, $C_d(u)$ will be approximately 0.50 when $u = d$. (A better, but still not exact, position is $u = d - \frac{2}{3}$.) (2) $C_d(u)$ will be approximately 0.15 when

$u = d + \sqrt{2d}$, and it will be approximately 0.025 when $u = d + 2\sqrt{2d}$. These approximations are useful for all values of d but are good to only one decimal place when $d \leq 3$. Note that for $d > 30$, these facts follow from the normal approximation given above with $z = 0, 1,$ and 2.

APPENDIX TO CHAPTER 11

This Appendix presents several further theoretical comments.

Theory of chi-square approximation. The paragraphs at the beginning of this chapter have given a procedure for using chi-square curves to get a descriptive level of significance when a single multinomial model is assumed. Why and how does this procedure work? First, we look at the form of the approximation. Then we look further at chi-square variables and degrees of free

We return to the Boston-Montreal example considered in Chapter 10 and at the beginning of our description of chi-square approximation. We noted, in Chapter 10, that there are 66 possible observations for this experiment. (Recall that our model has $p_1 = 0.5$, $p_2 = 0.3$, $p_3 = 0.2$). For each of the 66 observations, we can calculate (i) the value of the CS metric, and (ii) by the multinomial formula, the probability that the observation occurs. We can then make a list of the 66 possible observations in increasing order of CS value, and for each we can give its probability. This list would begin as follows (we give the first seven lines):

Observation	χ_o^2	P(Observation)
(5,3,2)	0.0	0.085
(4,4,2)	0.53	0.064
(6,2,2)	0.53	0.071
(6,3,1)	0.70	0.071
(4,3,3)	0.70	0.057
(5,4,1)	0.83	0.064
(5,2,3)	0.83	0.057
. . . .	. . . .	. . . .

We could now present this list graphically by taking χ_o^2 values on the horizontal axis, probability values on the vertical axis, and plotting a single point for each observation. This graph would begin:

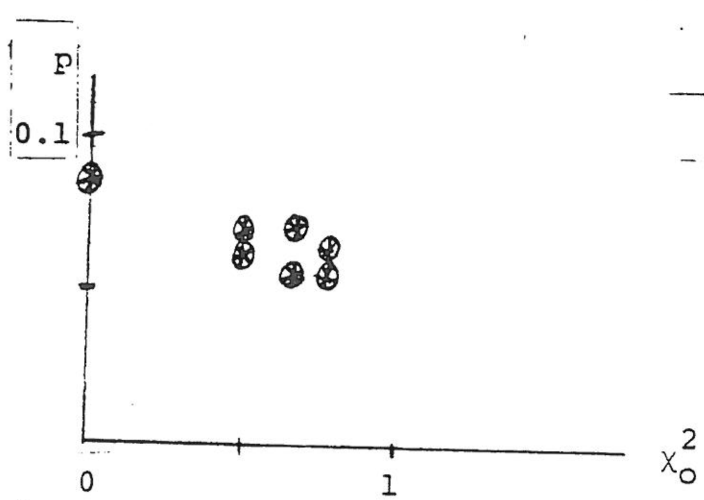

Figure 11.4

Note the irregular horizontal and vertical spacing of the points.

We now go on to give another, different, and more useful graphical picture of this list. We first break the list into 8 successive groups of approximately equal size (each group contains 7, 8, or 9 possible observations). (Note that for m = the total number of possible observations, we have divided the possible observations into approximately $\sqrt{m}$ groups, each of which contains approximately $\sqrt{m}$ observations.) We then list these groups, giving the interval of CS values for each group and the total probability for each group. (The interval is taken from the lowest value for each group to the lowest value for the next group.) Finally, we calculate the height that a bar would have to have on this interval in order that its area would be the total probability value for its group. Thus, in the following table, the final column is the quotient of the previous two columns.

Groups of possible observations	x_o^2 Interval	Interval length	Total probability	Bar height
First nine	[0,2.03)	2.03	0.552	0.27
Next eight	[2.03,3.33)	1.30	0.237	0.18
Next eight	[3.33,4.80)	1.47	0.120	0.08
Next nine	[4.80,6.70)	1.90	0.061	0.03
Next seven	[6.70,10.00)	3.30	0.019	0.01
Next eight Next eight Next nine	$\geq$ 10.00		0.01	0.00

Drawing the final bar graph, we get:

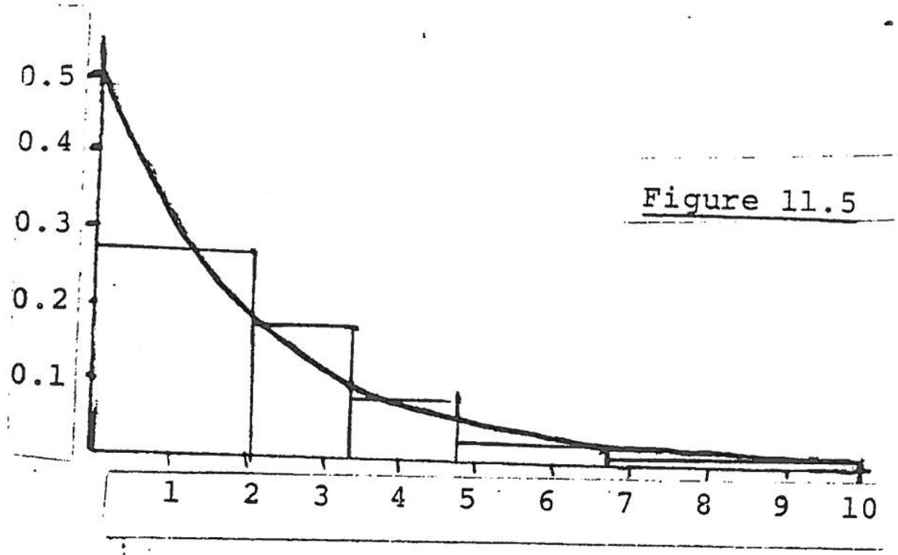

Figure 11.5

If we take the same multinomial probabilities, but let n, the total number of trials, increase; and if we use approximately $\sqrt{m}$ bars with the area of each bar representing the combined probability of approximately $\sqrt{m}$ possible observations contained in a corresponding interval of CS values (here, from the note at the end of the discussion of multinomial coefficients in Chapter 4, $m = \left(\binom{n+2}{n}\right)$; then the resulting bar graph will have an increasing number of bars and will take on an increasingly smooth shape. It can be proved, by methods similar to those used in proving the normal approximation for binomial experiments, that this limiting shape is exactly given by the chi-square curve with 2 degrees of freedom. (This curve is superimposed on the bar graph in Figure 11.5.) More generally, it can be proved that for a multinomial experiment with c categories, the graph constructed as above will approach the exact shape of the chi-square curve with $c - 1$ degrees of freedom, as n increases. It follows from this fact that the procedure of chi-square approximation described at the beginning of this chapter will give good approximations to the desired descriptive levels of significance, since a descriptive level of significance will be closely approximated by a corresponding area of the bar graph, and an area of the bar graph will be closely approximated by a corresponding area under the chi-square curve. In the next paragraphs, we comment further on the proof of chi-square approximation.

Chi-square variables. Let X be a random variable whose distribution is given by the standard normal curve. Consider the new random variable $Y = X^2$. Let $g(y)$ be the probability density for Y. Then we have

$$\int_{-\infty}^{u} g(y)\,dy = P(Y < u) = P(X^2 < u)$$

$$= P(|X| < \sqrt{u}) = 2\int_{0}^{\sqrt{u}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx .$$

Differentiating the two integrals with respect to u, we have, by the fundamental theorem of integral calculus, that

$$g(u) = \frac{1}{\sqrt{2\pi}} u^{-1/2} e^{-\frac{u}{2}} .$$

Thus the probability density for Y is given by the chi-square curve with one degree of freedom. Moreover, we can show, by integrating, that

$$E_Y = 1 \quad \text{and} \quad V_Y = 2 .$$

Let $X_1, X_2, \ldots, X_d$ be independent random variables, each with a standard normal distribution. Consider the new random variable $Z = X_1^2 + X_2^2 + \ldots + X_d^2$. Using the independence of $X_1^2, X_2^2, \ldots, X_d^2$ and using the formula obtained in (2) on pages 236-237 of Chapter 8 for the probability density of a sum of independent random variables, we can show that the distribution of Z is given by the chi-square curve with d degrees of freedom. Morever, it follows from rules (II) and (VII) in Chapter 8 that

$$E_Z = d \quad \text{and} \quad V_Z = 2d \; .$$

Degrees of freedom. Take the following experiment: do c independent binomial experiments, each with n individual trials, and with probabilities of individual success $p_1, p_2, \ldots, p_c$. Let $x_1, x_2, \ldots, x_c$ be the resulting observation. Form the metric:

$$N(x_1, \ldots, x_c) = \frac{(x_1 - np_1)^2}{np_1q_1} + \frac{(x_2 - np_2)^2}{np_2q_2} + \ldots + \frac{(x_c - np_c)^2}{np_cq_c} \; .$$

(Note that this is like the CS metric, except for the appearance of $q_1, \ldots, q_c$ in the denominators.) It follows from normal approximation and from our discussion above of chi-square variables that, as n increases, approximate descriptive levels of significance for this experiment and this metric will be given by the chi-square curve with c degrees of freedom (and will not depend on the values of $p_1, \ldots, p_c$).

In this example of c independent binomial experiments, the observed values $x_1, \ldots, x_c$ are independent of one another. In the case of chi-square approximation for multinomial experiments with c categories, the numbers observed in the different categories are no longer independent. (For example if $c = 3$, $n = 10$, and $x_1 = 7$, then x_2 and x_3 must both be ≤ 3.) It is now possible to show, using normal approximation, that if we have a multinomial experiment with c categories, then we can use chi-square approximation to get DLS values, provided that: (i) we use np_i rather than np_iq_i in the denominators of our metric (this results in the CS metric in place of the metric N above); and (ii) we use, for our approximation, a chi-square curve with c-1 rather than c degrees of freedom.

We illustrate this for the case $c = 2$. It is an algebraic identity that the CS metric for a multinomial experiment with 2 categories,

$$\frac{(x_1 - np_1)^2}{np_1} + \frac{(x_2 - np_2)^2}{np_2},$$

is identical with the metric

$$\frac{(x_1 - np_1)^2}{np_1(1 - p_1)}$$

for a single binomial experiment. (We remarked on this in Chapter 10.) A similar argument, using the fact that $p_1 + p_2 + \ldots + p_c = 1$ in a multinomial experiment of c categories, shows that a CS metric for c categories gives the same descriptive levels as the metric N above for an appropriate set of c - 1 independent binomial experiments. In this way, the reduction from c to c - 1 degrees of freedom arises and a proof of chi-square approximation for multinomial experiments is obtained.

We have already noted, from our discussion of chi-square variables above, that the standard normal curve and the chi-square curve with 1 degree of freedom are closely related. In particular, for any $u > 0$, we must have that

$$C_1(u) = 1 - 2A(\sqrt{u}) \ ,$$

where $A(\sqrt{u})$ is area under the standard normal curve between 0 and $\sqrt{u}$. Hence values for C_1 can be found from a table of values for A. This is sometimes a convenient way to find values for C_1.

<u>Theoretical note</u>. We now return to the procedure, described earlier in the chapter, of using the observation to indicate a model in a <u>CS</u> metric <u>DLS</u> calculation. In that procedure, we used one linear algebraic fact about the observation (its average value) in order to fit the Poisson distribution. We then subtracted one further degree of freedom in doing chi-square approximation.

This over-all form of procedure (of using the observation to indicate a single model from a set of all models of a given kind and then using chi-square approximation) can be applied to certain other sets of models (besides the set of all Poisson models). We see an example of this in Exercise 11-16 below. This over-all form of procedure can also be applied to certain other experiments (besides experiments involving repeated trials on an infinite discrete sample space). These other experiments include experiments which are like multinomial experiments (repeated trials on a finite sample space) except that the nature of the experiment may require that certain additional linear algebraic relationships hold (for all observations) among the observed values X_1, X_2, ..., X_c. We see examples of this in Chapters 12 and 13.

In each such case, when we use an observation to indicate a single model from a given set of models and then apply chi-square approximation, we must adjust the degrees of freedom in the final chi-square approximation by subtracting one additional degree of freedom for each separate and independent linear numerical fact (about the specific observation, x_1, x_2, ...) which is used to help indicate a single model in the given set of models; and by subtracting one additional degree of freedom for each separate and independent linear algebraic relationship that is required to hold (for all observations) among X_1, X_2, ...). We shall later see that for such an over-all procedure to be valid, the numerical facts about a specific observation x_1, x_2, ... and the algebraic relationships among X_1, X_2, ... (for all observations) must be expressible by linear equations.

In each case, we shall find that an appropriate version of chi-square approximation gives a DLS value (for some suitable metric like the metric s above) which is approximately the same for all models in the given set. In later chapters we shall, from time to time, return to consider the concepts upon which this extraordinary result is based, and we shall eventually provide the outline of a proof for it.

CHI-SQUARE TABLES (inverse form)

(Giving u as a function of degrees of freedom (on the left) and Area C_d (across the top).)

Degrees of freedom \ Area C_d:	0.99	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.01	0.001
1	0.00	0.00	0.02	0.06	0.15	0.27	0.45	0.71	1.07	1.64	2.71	3.84	6.63	10.83
2	0.02	0.10	0.21	0.45	0.71	1.02	1.39	1.83	2.41	3.22	4.61	5.99	9.21	13.82
3	0.11	0.35	0.58	1.01	1.42	1.87	2.37	2.95	3.66	4.64	6.25	7.81	11.34	16.27
4	0.30	0.71	1.06	1.65	2.19	2.75	3.36	4.04	4.88	5.99	7.78	9.49	13.28	18.47
5	0.55	1.15	1.61	2.34	3.00	3.66	4.35	5.13	6.06	7.29	9.24	11.07	15.09	20.52
6	0.87	1.64	2.20	3.07	3.83	4.57	5.35	6.21	7.23	8.56	10.64	12.59	16.81	22.46
7	1.24	2.17	2.83	3.82	4.67	5.49	6.35	7.28	8.38	9.80	12.02	14.07	18.48	24.32
8	1.65	2.73	3.49	4.59	5.53	6.42	7.34	8.35	9.52	11.03	13.36	15.51	20.09	26.12
9	2.09	3.33	4.17	5.38	6.39	7.36	8.34	9.41	10.66	12.24	14.68	16.92	21.67	27.88
10	2.56	3.94	4.87	6.18	7.27	8.30	9.34	10.47	11.73	13.44	15.99	18.31	23.21	29.59
11	3.05	4.57	5.58	6.99	8.15	9.24	10.34	11.53	12.90	14.63	17.28	19.68	24.72	31.26
12	3.57	5.23	6.30	7.81	9.03	10.18	11.34	12.58	14.01	15.81	18.55	21.03	26.22	32.91
13	4.11	5.89	7.04	8.63	9.93	11.13	12.34	13.64	15.12	16.98	19.81	22.36	27.69	34.53
14	4.66	6.57	7.79	9.47	10.82	12.08	13.34	14.69	16.22	18.15	21.06	23.68	29.14	36.12
15	5.23	7.26	8.55	10.31	11.72	13.03	14.34	15.73	17.32	19.31	22.31	25.00	30.58	37.70
16	5.81	7.96	9.31	11.15	12.62	13.98	15.34	16.78	18.42	20.47	23.54	26.30	32.00	39.25
17	6.41	8.67	10.09	12.00	13.53	14.94	16.34	17.82	19.51	21.61	24.77	27.59	33.41	40.79
18	7.01	9.39	10.86	12.86	14.44	15.89	17.34	18.87	20.60	22.76	25.99	28.87	34.81	42.31
19	7.63	10.12	11.65	13.72	15.35	16.85	18.34	19.91	21.69	23.90	27.20	30.14	36.19	43.82
20	8.26	10.85	12.44	14.58	16.27	17.81	19.34	20.95	22.77	25.04	28.41	31.41	37.57	45.31
21	8.90	11.59	13.24	15.44	17.18	18.77	20.34	21.99	23.86	26.17	29.62	32.67	38.93	46.80
22	9.54	12.34	14.04	16.31	18.10	19.73	21.34	23.03	24.94	27.30	30.81	33.92	40.29	48.27
23	10.20	13.09	14.85	17.19	19.02	20.69	22.34	24.07	26.02	28.43	32.01	35.17	41.64	49.73
24	10.86	13.85	15.66	18.06	19.94	21.65	23.34	25.11	27.10	29.55	33.20	36.42	42.98	51.18
25	11.52	14.61	16.47	18.94	20.87	22.62	24.34	26.14	28.17	30.68	34.38	37.65	44.31	52.62
26	12.20	15.38	17.29	19.82	21.79	23.58	25.34	27.18	29.25	31.79	35.56	38.89	45.64	54.05
27	12.88	16.15	18.11	20.70	22.72	24.54	26.34	28.21	30.32	32.91	36.74	40.11	46.96	55.48
28	13.56	16.93	18.94	21.59	23.65	25.51	27.34	29.25	31.39	34.03	37.92	41.34	48.28	56.89
29	14.26	17.71	19.77	22.48	24.58	26.48	28.34	30.28	32.46	35.14	39.09	42.56	49.59	58.30
30	14.95	18.49	20.60	23.36	25.51	27.44	29.34	31.32	33.53	36.25	40.26	43.77	50.89	59.70
50	29.71	34.76	37.69	41.45	44.31	46.86	49.33	51.89	54.72	58.16	63.17	67.50	76.15	86.66
100	70.06	77.93	82.36	87.95	92.13	95.81	99.33	102.9	106.9	111.7	118.5	124.3	135.8	149.4
200	156.4	168.3	174.8	183.0	189.0	194.3	199.3	204.4	210.0	216.6	226.0	234.0	249.4	267.5
500	429.4	449.1	459.9	473.2	482.9	491.4	499.3	507.4	516.1	526.4	540.9	553.1	576.5	603.4
800	709.9	735.4	749.2	766.2	778.6	789.2	799.3	809.5	820.5	833.5	851.7	866.9	896.0	929.3
1000	898.9	927.6	943.1	962.2	976.1	988.4	999.3	1011.	1023.	1037.	1058.	1075.	1107.	1144.

EXERCISES FOR CHAPTER 11.

Note. In the following exercises, unless it is stated otherwise, chi-square approximation should be used to find DLS values, even though the full conditions for two-decimal-place accuracy may not apply. Pooling should be done as necessary. Keep in mind that pooling is done on the basis of expected values, not observed values.

11-1. A die is rolled 36 times. 7 ones, 5 twos, 4 threes, 9 fours, 3 fives, and 8 sixes are obtained. Estimate the DLS of this observation under the model for a fair die.

11-2. In 20 games against Montreal, the Boston hockey team wins 12, loses 4, and ties 4. Assume, as a model, that the probability of winning is 0.4, of losing is 0.4, and of tying is 0.2, and that the individual outcomes are independent from game to game. What is the DLS of the observed result under this model?

11-3. A baseball player gets 25 single-base hits and 5 extra-base hits in 100 times at bat. Given, as a model, that the probability of a single-base hit is 0.2 in a time at bat, that the probability of an extra-base hit is 0.1, and that the individual outcome is independent from time at bat to time at bat, find the DLS of the batter's observed performance. (The value of the CS metric for this example was calculated in Exercise 10-9.)

11-4. (a) On page 259, data were given for occurrences of different digits in a certain observed sequence of 100 pseudorandom digits. Find the <u>DLS</u> of these data under the model that the ten digits occur independently and with equal probabilities.

(b) On page 259, data were given for occurrences of different digits in a certain observed sequence of 100 random digits. Find the <u>DLS</u> of these data under the model that the digits are in fact random.

11-5. Use random digits to simulate a multinomial experiment with $c = 3$, $n = 50$, $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$. Find a <u>DLS</u> for the observation you obtain. (<u>Suggestion</u>. Use line 40 on page 274. Let 1, 2, 3, 4, 5 represent the first category, 6, 7, 8 the second, and 9, 0 the third.)

11-6. In 50 blocks of 5 random digits each, observe the number of <u>zeros</u> that occur in each block. Make a table showing the number of blocks with x zeros, for $x = 0, 1, 2, 3, 4, 5$. Treat this as a multinomial experiment with $c = 6$ and $n = 50$. Assuming that the digits are in fact random, find a <u>DLS</u> for your observation. (<u>Suggestion</u>. Use lines 41 through 46 on page 274.)

11-7. For a certain experiment in the cross-breeding of plants, genetic theory asserts that two kinds of offspring, one with physical characteristic A and one with physical characteristic a, will occur with probabilities 3/4 and 1/4 respectively. (<u>Note</u>. This model holds, for example, for the second generation offspring of two pure-bred parents of kinds A and a, where characteristic A is the <u>dominant</u> form of a simple Mendelian trait and a is the corresponding <u>recessive</u> form.) The experiment is carried out, and the following counts are observed:

A	a
112	30

(This result is from an experiment by Mendel.) Find the <u>DLS</u> of this observation under the given theoretical model.

11-8. For a certain experiment in the cross-breeding of plants, genetic theory asserts that four kinds of offspring, one with physical characteristics A and B, one with A and b, one with a and B, and one with a and b, will occur with probabilities 9/16, 3/16, 3/16, and 1/16 respectively. (<u>Note</u>. This model holds, for example, for the second generation offspring of two pure-bred parents of kinds AB and ab, where A and B are the <u>dominant</u> forms of two independent simple Mendelian traits, and where a and b are the corresponding <u>recessive</u> forms. Some of Mendel's original experiments with garden peas were of this kind.)

The experiment is carried out, and the following counts are observed:

AB	Ab	aB	ab
150	62	51	25.

(These data are from an experiment by Mendel.) Find the <u>DLS</u> of this observation under the given theoretical model.

11-9. Forty blocks of 50 random digits each are observed (the first 40 lines on page 231), and the number of <u>zeros</u> in each block is counted. Let x be the number of <u>zeros</u> in a block. The observed distribution of x is found to be:

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	≥ 14
n_x	0	0	2	5	7	11	8	3	1	0	1	1	0	1	0

where n_x is the number of blocks with exactly x <u>zeros</u>. Take, as a theoretical distribution, the binomial distribution with $n = 50$ and $p = 0.1$. Find the goodness-of-fit (in terms of a <u>DLS</u> value) of the observed distribution. (Poisson approximation may be used. Pool as necessary.)

11-10. Each trial of a certain experiment produces a real number between 0 and 1. (The physical experiment might be to spin a horizontal pointer, pivoted at its center, and to observe, when the pointer comes to rest, the position of the point on a circular scale that ranges from 0 to 1 in one full turn.) In thirty trials, the following values are observed (to two-decimal-place accuracy):

0.44, 0.31, 0.92, 0.23, 0.13, 0.64, 0.74, 0.15, 0.21, 0.06, 0.28, 0.46, 0.78, 0.05, 0.47, 0.76, 0.99, 0.34 0.51, 0.20, 0.87, 0.28, 0.22, 0.17, 0.25, 0.03, 0.30, 0.32, 0.57, 0.50, 0.30, 0.87, 0.42, 0.65, 0.05, 0.26, 0.19, 0.27, 0.42, 0.86, 0.04, 0.06, 0.33, 0.28, 0.24, 0.56, 0.05, 0.00, 0.34, 0.64.

Assume, as model, a continuous probability space with the interval $0 \leq x < 1$ as the sample space and with the constant function $f(x) = 1$ as the probability density function. Find the goodness-of-fit (in terms of a DLS value) of the observed data to the assumed model. (Hint. Use the sub-intervals (0.00, 0.09), (0.10, 0.19),...,(0.90,0.99), and treat the experiment as a multinomial experiment with $c = 10$ and $n = 50$.)

11-11. The first six lines in the table of random digits on page 272 are examined, and the sizes of the intervals between successive occurrences of zero are recorded, where the size of an interval is the number of non-zero digits occurring in that interval. We obtain the following data:

22, 7, 10, 7, 0, 1, 1, 4, 1, 3, 9, 0, 2, 7, 3, 22, 4, 1, 3, 3, 21, 9, 6, 0, 6, 12, 10, 6, 14, 5, 12, 15, 2, 12, 10, 12, 1.

(a) Give an expression for the probability that an interval of size x occurs, assuming that the digits are truly random.

(b) Find the goodness-of-fit of the observed data to this model, in terms of an appropriate DLS value. (Hint. Use a calculator to find the probability values and the expected values. Pool as necessary.)

11-12. Figures for the number of goals per period in 31 National Hockey League games played between February 25 and March 2, 1980 were as follows:

x	0	1	2	3	4	5	≥ 6
O_x	12	19	27	15	15	5	0

Here O_x is the number of periods observed in which the goal total was x.

(a) Take the Poisson distribution with $m = 2.35$ as a given model. Find the DLS of these data under this given model (as a measure of goodness-of-fit of the observed distribution to the assumed Poisson distribution.)

(b) Find a DLS under the assumption that the model is some (otherwise unspecified) Poisson distribution. (Use the observation to indicate a specific model by fitting a Poisson distribution to the data.)

(c) Explain why the DLS value for (b) is smaller than the value for (a), even though the Poisson distribution indicated and used in (b) fits the data better than the Poisson distribution assumed in (a).

11-13. Take the data on deaths from horse-kicks on page 184.

(a) Find the DLS of these data under the assumption that the model is some (otherwise unspecified) Poisson distribution.

(b) Assume that the Poisson distribution $m = 0.61$ has in fact been suggested ahead of time, before you observe the data, and that you then observe the given data. What is the DLS of the data under this given model?

(c) Explain why the DLS value found in (b) is larger than the DLS value found in (a). In particular, carefully describe an experiment for which the DLS value found in (a) is a correct probability value and an experiment for which the DLS value found in (b) is a correct probability value.

11-14. Take the data on flying-bomb hits on page 185. Find the DLS of these data under the assumption that the model is some (otherwise unspecified) Poisson distribution.

11-15. In a certain telephone exchange, the intervals between the beginnings of successive telephone calls are measured to the nearest tenth of a second. Twenty intervals are observed, and the following measurements are recorded: 0.1, 0.5, 0.1, 0.9, 1.2, 0.1, 0.3, 0.3, 0.2, 1.0, 0.8, 0.2, 0.6, 0.1, 0.7, 0.6, 0.4, 0.4, 0.7, 0.2. We wish to calculate a DLS for this observation on the assumption that the correct model is a continuous probability space with a probability density function of the form me^{-mx} on the interval $0 < x < \infty$. (See pages 54-56.) Assume that the data may be used to indicate a model for a CS metric DLS calculation.

(a) Use these data to indicate a model of the desired kind. (Hint. Note that the expected average interval length under the model me^{-mx} will be

$$\int_0^\infty xme^{-mx}\,dx = \frac{1}{m}.)$$

(b) Use the above data with this indicated model to get a DLS value by chi-square approximation. (Hint. Pool into four subintervals, each with probability 0.25.)

11-16. In a certain country, each day is either rainy or sunny. A sequence of rainy and sunny days is observed and the sizes of the intervals between successive sunny days is recorded as in the following table:

Interval size	0	1	2	3	4	5	6	≥ 7
Number of intervals	22	7	5	10	8	7	1	0

Here the size of an interval is the number of rainy days in that interval. We wish to calculate a DLS for this observation on the assumption that the correct model is a sequence of independent Bernoulli trials with some (unspecified) value for p, the probability of a sunny day on a single trial. Assume that the average observed interval length may be used to indicate a model for the purpose of a CS-metric-DLS calculation.

(a) Use the data to indicate a particular Bernoulli model (that is to say, a particular value of p). (Hint. See Exercise 9-9.)

(b) Use the above data, with this model, to get a DLS value by chi-square approximation.

11-17. A multinomial experiment with 6 categories and $n = 60$ yields the observation (6,14,16,4,10,10).

(a) Assume that the model is some unspecified model from the set of multinomial models (with $c = 6$ and $n = 60$) satisfying the special condition that $p_1 = p_2 = p_3$ and $p_4 = p_5 = p_6$. Use the observed average of x_1, x_2, x_3 to indicate a model, and find a DLS by chi-square approximation. (See the discussion of degrees of freedom in the Appendix above.)

(b) Assume that the model is some unspecified model from the set of multinomial models (with $c = 6$ and $n = 60$) satisfying the special condition that $p_1 = p_2$, $p_3 = p_4$, and $p_5 = p_6$. Use the observed averages of x_1, x_2 and of x_3, x_4 to indicate a model, and find a DLS by chi-square approximation. (See the Appendix.)

(c) In each of the above cases, carefully describe the metric and the experiment to which the final DLS value refers.

11-18. (a) In a particular sequence of 380 plays at roulette, the 38 different pocket-numbers are observed to occur as shown in the following table:

Pocket-number	0	00	1	2	3	4	5	6	7
Occurrences	10	8	7	12	6	10	4	9	13

8	9	10	11	12	13	14	15	16	17	18	19
9	8	14	8	15	9	11	9	13	6	8	11

20	21	22	23	24	25	26	27	28	29
9	7	9	8	11	11	13	10	17	11

30	31	32	33	34	35	36
10	11	15	6	11	8	13

Consider this experiment as a multinomial experiment with $c = 38$ and $n = 380$. Find the DLS of the observed result assuming that the wheel is fair.

(b) The data in (a) can also be given in the form of the following table:

f	4	5	6	7	8	9	10	11	12	13	14	15	16	17	≥ 18
n_f	1	0	3	2	6	6	4	7	1	4	1	2	0	1	0

where, for each f, n_f is the number of pocket-numbers occurring exactly f times. Note that $\sum_f n_f = 38$ and that $\sum_f f n_f = 380$.

Assume as a model for the distribution in this table a Poisson distribution with $m = 10$. Find, as a DLS value, the goodness-of-fit of the observed table to the

Poisson distribution. (<u>Note</u>. The Poisson distribution is a reasonable model, since observation of the number of occurrences of any single pocket-number in 380 plays can be viewed as a binomial experiment with $p = 1/38$ and $n = 380$, and since the observed values for different pocket-numbers are <u>nearly</u> independent (in a sense which we do not make precise here). The Poisson model then follows by Poisson approximation. The Poisson model cannot be fully justified, however, since the observation for one pocket-number is not fully independent of the observation of another pocket-number in the same sequence of 380 plays.)

<u>Comment</u>. Which <u>DLS</u> value is correct in Exercise 11-18, the one found in (a) and based on a multinomial experiment, or the one found in (b) and based on goodness-of-fit to a Poisson distribution? The two values come, in reality, from the use of two different metrics. The better choice of a metric in this case depends, as we shall later see, on what we conceive to be the possible alternatives to the assumed underlying model. In the present case, if we know that the experiment is <u>necessarily multinomial</u>, then approach (a) is better. On the other hand, if we suspect that the operator of the wheel may be <u>secretly controlling each outcome</u>, then approach (b) may be better. We shall consider these matters further in Chapters 12 and 14.