GOODNESS OF FIT TESTS AND CATEGORICAL DATA ANALYSIS

11.1 INTRODUCTION

We are often interested in determining whether or not a particular probabilistic model is appropriate for a given random phenomenon. This determination often reduces to testing whether a given random sample comes from some specified, or partially specified, probability distribution. For example, we may *a priori* feel that the number of industrial accidents occurring daily at a particular plant should constitute a random sample from a Poisson distribution. This hypothesis can then be tested by observing the number of accidents over a sequence of days and then testing whether it is reasonable to suppose that the underlying distribution is Poisson. Statistical tests that determine whether a given probabilistic mechanism is appropriate are called *goodness of fit* tests.

The classical approach to obtaining a goodness of fit test of a null hypothesis that a sample has a specified probability distribution is to partition the possible values of the random variables into a finite number of regions. The numbers of the sample values that fall within each region are then determined and compared with the theoretical expected numbers under the specified probability distribution, and when they are significantly different the null hypothesis is rejected. The details of such a test are presented in Section 11.2, where it is assumed that the null hypothesis probability distribution is completely specified. In Section 11.3, we show how to do the analysis when some of the parameters of the null hypothesis distribution are left unspecified; that is, for instance, the null hypothesis might be that the sample distribution is a normal distribution, without specifying the mean and variance of this distribution. In Sections 11.4 and 11.5, we consider situations where each member of a population is classified according to two distinct characteristics, and we show how to use our previous analysis to test the hypothesis that the characteristics of a randomly chosen member of the population are independent. As an application, we show how to test the hypothesis that m population all have the same discrete probability distribution. Finally, in the optional section, Section 11.6, we return

to the problem of testing that sample data come from a specified probability distribution, which we now assume is continuous. Rather than discretizing the data so as to be able to use the test of Section 11.2, we treat the data as given and make use of the *Kolmogorov–Smirnov test*.

11.2 GOODNESS OF FIT TESTS WHEN ALL PARAMETERS ARE SPECIFIED

Suppose that n independent random variables — $Y_1, \ldots, Y_n$, each taking on one of the values $1, 2, \ldots, k$ — are to be observed and we are interested in testing the null hypothesis that $\{p_i, i = 1, \ldots, k\}$ is the probability mass function of the Y_j. That is, if Y represents any of the Y_j, then the null hypothesis is

$$H_0 : P\{Y = i\} = p_i, \qquad i = 1, \ldots, k$$

whereas the alternative hypothesis is

$$H_1 : P\{Y = i\} \neq p_i, \qquad \text{for some } i = 1, \ldots, k$$

To test the foregoing hypothesis, let $X_i, i = 1, \ldots, k$, denote the number of the Y_j's that equal i. Then as each Y_j will independently equal i with probability $P\{Y = i\}$, it follows that, under H_0, X_i is binomial with parameters n and p_i. Hence, when H_0 is true,

$$E[X_i] = np_i$$

and so $(X_i - np_i)^2$ will be an indication as to how likely it appears that p_i indeed equals the probability that $Y = i$. When this is large, say, in relationship to np_i, then it is an indication that H_0 is not correct. Indeed such reasoning leads us to consider the following test statistic:

$$T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{11.2.1}$$

and to reject the null hypothesis when T is large.

To determine the critical region, we need first specify a significance level α and then we must determine that critical value c such that

$$P_{H_0}\{T \geq c\} = \alpha$$

That is, we need to determine c so that the probability that the test statistic T is at least as large as c, when H_0 is true, is α. The test is then to reject the hypothesis, at the α level of significance, when $T \geq c$ and to accept when $T < c$.

It remains to determine c. The classical approach to doing so is to use the result that when n is large T will have, when H_0 is true, approximately (with the approximation

becoming exact as n approaches infinity) a chi-square distribution with $k-1$ degrees of freedom. Hence, for n large, c can be taken to equal $\chi^2_{\alpha,k-1}$; and so the approximate α level test is

$$\begin{aligned} &\text{reject} \quad H_0 \quad \text{if} \quad T \geq \chi^2_{\alpha,k-1} \\ &\text{accept} \quad H_0 \quad \text{otherwise} \end{aligned}$$

If the observed value of T is $T = t$, then the preceding test is equivalent to rejecting H_0 if the significance level α is at least as large as the p-value given by

$$\begin{aligned} p\text{-value} &= P_{H_0}\{T \geq t\} \\ &\approx P\{\chi^2_{k-1} \geq t\} \end{aligned}$$

where χ^2_{k-1} is a chi-square random variable with $k-1$ degrees of freedom.

An accepted rule of thumb as to how large n need be for the foregoing to be a good approximation is that it should be large enough so that $np_i \geq 1$ for each $i, i = 1, \ldots, k$, and also at least 80 percent of the values np_i should exceed 5.

REMARKS

(a) A computationally simpler formula for T can be obtained by expanding the square in Equation 11.2.1 and using the results that $\sum_i p_i = 1$ and $\sum_i X_i = n$ (why is this true?):

$$\begin{aligned} T &= \sum_{i=1}^{k} \frac{X_i^2 - 2np_iX_i + n^2p_i^2}{np_i} \qquad (11.2.2) \\ &= \sum_i X_i^2/np_i - 2\sum_i X_i + n\sum_i p_i \\ &= \sum_i X_i^2/np_i - n \end{aligned}$$

(b) The intuitive reason why T, which depends on the k values $X_1, \ldots, X_k$, has only $k-1$ degrees of freedom is that 1 degree of freedom is lost because of the linear relationship $\sum_i X_i = n$.

(c) Whereas the proof that, asymptotically, T has a chi-square distribution is advanced, it can be easily shown when $k = 2$. In this case, since $X_1 + X_2 = n$, and $p_1 + p_2 = 1$, we

see that

$$\begin{aligned} T &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_1 - n[1 - p_1])^2}{n(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} \qquad \text{since} \qquad \frac{1}{p} + \frac{1}{1 - p} = \frac{1}{p(1 - p)} \end{aligned}$$

However, X_1 is a binomial random variable with mean np_1 and variance $np_1(1 - p_1)$ and thus, by the normal approximation to the binomial, it follows that $(X_1 - np_1)/\sqrt{np_1(1 - p_1)}$ has, for large n, approximately a standard normal distribution, and so its square has approximately a chi-square distribution with 1 degree of freedom.

EXAMPLE 11.2a In recent years, a correlation between mental and physical well-being has increasingly become accepted. An analysis of birthdays and death days of famous people could be used as further evidence in the study of this correlation. To use these data, we are supposing that being able to look forward to something betters a person's mental state, and that a famous person would probably look forward to his or her birthday because of the resulting attention, affection, and so on. If a famous person is in poor health and dying, then perhaps anticipating his birthday would "cheer him up and therefore improve his health and possibly decrease the chance that he will die shortly before his birthday." The data might therefore reveal that a famous person is less likely to die in the months before his or her birthday and more likely to die in the months afterward.

SOLUTION To test this, a sample of 1,251 (deceased) Americans was randomly chosen from *Who Was Who in America*, and their birth and death days were noted. (The data are taken from D. Phillips, "Death Day and Birthday: An Unexpected Connection," in *Statistics: A Guide to the Unknown*, Holden-Day, 1972.) The data are summarized in Table 11.1.

If the death day does not depend on the birthday, then it would seem that each of the 1,251 individuals would be equally likely to fall in any of the 12 categories. Thus, let us test the null hypothesis

$$H_0 = p_i = \frac{1}{12}, \qquad i = 1, \ldots, 12$$

TABLE 11.1 *Number of Deaths Before, During, and After the Birth Month*

	6 Months Before	5 Months Before	4 Months Before	3 Months Before	2 Months Before	1 Month Before	The Month	1 Month After	2 Months After	3 Months After	4 Months After	5 Months After
Number of deaths	90	100	87	96	101	86	119	118	121	114	113	106

$n = 1{,}251$
$n/12 = 104.25$

Since $np_i = 1,251/12 = 104.25$, the chi-square test statistic for this hypothesis is

$$\begin{aligned} T &= \frac{(90)^2 + (100)^2 + (87)^2 + \cdots + (106)^2}{104.25} - 1{,}251 \\ &= 17.192 \end{aligned}$$

The p-value is

$$\begin{aligned} p\text{-value} &\approx P\{\chi^2_{11} \geq 17.192\} \\ &= 1 - .8977 = .1023 \quad \text{by Program 5.8.1a} \end{aligned}$$

The results of this test leave us somewhat up in the air about the hypothesis that an approaching birthday has no effect on an individual's remaining lifetime. For whereas the data are not quite strong enough (at least, at the 10 percent level of significance) to reject this hypothesis, they are certainly suggestive of its possible falsity. This raises the possibility that perhaps we should not have allowed as many as 12 data categories, and that we might have obtained a more powerful test by allowing for a fewer number of possible outcomes. For instance, let us determine what the result would have been if we had coded the data into 4 possible outcomes as follows:

$$\begin{aligned} &\text{outcome } 1 = -6, -5, -4 \\ &\text{outcome } 2 = -3, -2, -1 \\ &\text{outcome } 3 = 0, 1, 2 \\ &\text{outcome } 4 = 3, 4, 5 \end{aligned}$$

That is, for instance, an individual whose death day occurred 3 months before his or her birthday would be placed in outcome 2. With this classification, the data would be as follows:

Outcome	Number of Times Occurring
1	277
2	283
3	358
4	333

$n = 1{,}251$
$n/4 = 312.75$

The test statistic for testing $H_0 = p_i = 1/4, i = 1, 2, 3, 4$ is

$$\begin{aligned} T &= \frac{(277)^2 + (283)^2 + (358)^2 + (333)^2}{312.75} - 1.251 \\ &= 14.775 \end{aligned}$$

Hence, as $\chi^2_{.01,3} = 11.345$, the null hypothesis would be rejected even at the 1 percent level of significance. Indeed, using Program 5.8.1a yields that

$$p\text{-value} \approx P\{\chi^2_3 \geq 14.775\} = 1 - .998 = .002$$

The foregoing analysis is, however, subject to the criticism that the null hypothesis was chosen after the data were observed. Indeed, while there is nothing incorrect about using a set of data to determine the "correct way" of phrasing a null hypothesis, the additional use of those data to test that very hypothesis is certainly questionable. Therefore, to be quite certain of the conclusion to be drawn from this example, it seems prudent to choose a second random sample — coding the values as before — and again test $H_0 : p_i = 1/4, i = 1, 2, 3, 4$ (see Problem 3). ■

Program 11.2.1 can be used to quickly calculate the value of T.

EXAMPLE 11.2b A contractor who purchases a large number of fluorescent lightbulbs has been told by the manufacturer that these bulbs are not of uniform quality but rather have been produced in such a way that each bulb produced will, independently, either be of quality level A, B, C, D, or E, with respective probabilities .15, .25, .35, .20, .05. However, the contractor feels that he is receiving too many type E (the lowest quality) bulbs, and so he decides to test the producer's claim by taking the time and expense to ascertain the quality of 30 such bulbs. Suppose that he discovers that of the 30 bulbs, 3 are of quality level A, 6 are of quality level B, 9 are of quality level C, 7 are of quality level D, and 5 are of quality level E. Do these data, at the 5 percent level of significance, enable the contractor to reject the producer's claim?

SOLUTION Program 11.2.1 gives the value of the test statistic as 9.348. Therefore,

$$\begin{aligned} p\text{-value} &= P_{H_0}\{T \geq 9.348\} \\ &\approx P\{\chi^2_4 \geq 9.348\} \\ &= 1 - .947 \quad \text{from Program 5.8.1a} \\ &= .053 \end{aligned}$$

Thus the hypothesis would not be rejected at the 5 percent level of significance (but since it would be rejected at all significance levels above .053, the contractor should certainly remain skeptical). ■

11.2.1 Determining the Critical Region by Simulation

From 1900 when Karl Pearson first showed that T has approximately (becoming exact as n approaches infinity) a chi-square distribution with $k - 1$ degrees of freedom, until relatively recently, this approximation was the only means available for determining the p-value of the goodness of fit test. However, with the advent of inexpensive, fast, and easily available computational power a second, potentially more accurate, approach has become available: namely, the use of simulation to obtain to a high level of accuracy the p-value of the test statistic.

The simulation approach is as follows. First, the value of T is determined — say, $T = t$. Now to determine whether or not to accept H_0, at a given significance level α, we need to know the probability that T would be at least as large as t when H_0 is true. To determine this probability, we simulate n independent random variables $Y_1^{(1)}, \ldots, Y_n^{(1)}$ each having the probability mass function $\{p_i, i = 1, \ldots, k\}$ — that is,

$$P\{Y_j^{(1)} = i\} = p_i, \qquad i = 1, \ldots, k, \qquad j = 1, \ldots, n$$

Now let

$$X_i^{(1)} = \text{number } j : Y_j^{(1)} = i$$

and set

$$T^{(1)} = \sum_{i=1}^{k} \frac{(X_i^{(1)} - np_i)^2}{np_i}$$

Now repeat this procedure by simulating a second set, independent of the first set, of n independent random variables $Y_1^{(2)}, \ldots, Y_n^{(2)}$ each having the probability mass function $\{p_i, i = 1, \ldots, k\}$ and then, as for the first set, determining $T^{(2)}$. Repeating this a large number, say, r, of times yields r independent random variables $T^{(1)}, T^{(2)}, \ldots, T^{(r)}$, each of which has the same distribution as does the test statistic T when H_0 is true. Hence, by the law of large numbers, the proportion of the T_i that are as large as t will be very nearly equal to the probability that T is as large as t when H_0 is true — that is,

$$\frac{\text{number } l : T^{(l)} \geq t}{r} \approx P_{H_0}\{T \geq t\}$$

In fact, by letting r be large, the foregoing can be considered to be, with high probability, almost an equality. Hence, if that proportion is less than or equal to α, then the p-value, equal to the probability of observing a T as large as t when H_0 is true, is less than α and so H_0 should be rejected.

REMARKS

(a) To utilize the foregoing simulation approach to determine whether or not to accept H_0 when T is observed, we need to specify how one can simulate, or generate, a random variable Y such that $P\{Y = i\} = p_i, i = 1, \ldots, k$. One way is as follows:

Step 1: Generate a random number U.
Step 2: If

$$p_1 + \cdots + p_{i-1} \le U < p_1 + \cdots + p_i$$

set $Y = i$ (where $p_1 + \cdots + p_{i-1} \equiv 0$ when $i = 1$). That is,

$$\begin{aligned} U < p_1 &\Rightarrow Y = 1 \\ p_1 \le U < p_1 + p_2 &\Rightarrow Y = 2 \\ &\vdots \\ p_1 + \cdots + p_{i-1} \le U < p_1 + \cdots + p_i &\Rightarrow Y = i \\ &\vdots \\ p_1 + \cdots + p_{n-1} < U &\Rightarrow Y = n \end{aligned}$$

Since a random number is equivalent to a uniform (0, 1) random variable, we have that

$$P\{a < U < b\} = b - a, \qquad 0 < a < b < 1$$

and so

$$P\{Y = i\} = P\{p_1 + \cdots + p_{i-1} < U < p_1 + \cdots + p_i\} = p_i$$

(**b**) A significant question that remains is how many simulation runs are necessary. It has been shown that the value $r = 100$ is usually sufficient at the conventional 5 percent level of significance.*

EXAMPLE 11.2c Let us reconsider the problem presented in Example 11.2b. A simulation study yielded the result

$$P_{H_0}\{T \le 9.52381\} = .95$$

and so the critical value should be 9.52381, which is remarkably close to $\chi^2_{.05,4} = 9.488$ given as the critical value by the chi-square approximation. This is most interesting since the rule of thumb for when the chi-square approximation can be applied — namely, that

* See Hope, A., "A Simplified Monte Carlo Significance Test Procedure," *J. of Royal Statist. Soc.*, B. 30, 582–598, 1968.

each $np_i \geq 1$ and at least 80 percent of the np_i exceed 5 — does not apply, thus raising the possibility that it is rather conservative. ■

Program 11.2.2 can be utilized to determine the p-value.

To obtain more information as to how well the chi-square approximation performs, consider the following example.

EXAMPLE 11.2d Consider an experiment having six possible outcomes whose probabilities are hypothesized to be .1, .1, .05, .4, .2, and .15. This is to be tested by performing 40 independent replications of the experiment. If the resultant number of times that each of the six outcomes occurs is 3, 3, 5, 18, 4, 7, should the hypothesis be accepted?

SOLUTION A direct computation, or the use of Program 11.2.1, yields that the value of the test statistic is 7.4167. Utilizing Program 5.8.1a gives the result that

$$P\{\chi_5^2 \leq 7.4167\} = .8088$$

and so

$$p\text{-value} \approx .1912$$

To check the foregoing approximation, we ran Program 11.2.2, using 10,000 simulation runs, and obtained an estimate of the p-value equal to .1843 (see Figure 11.1).

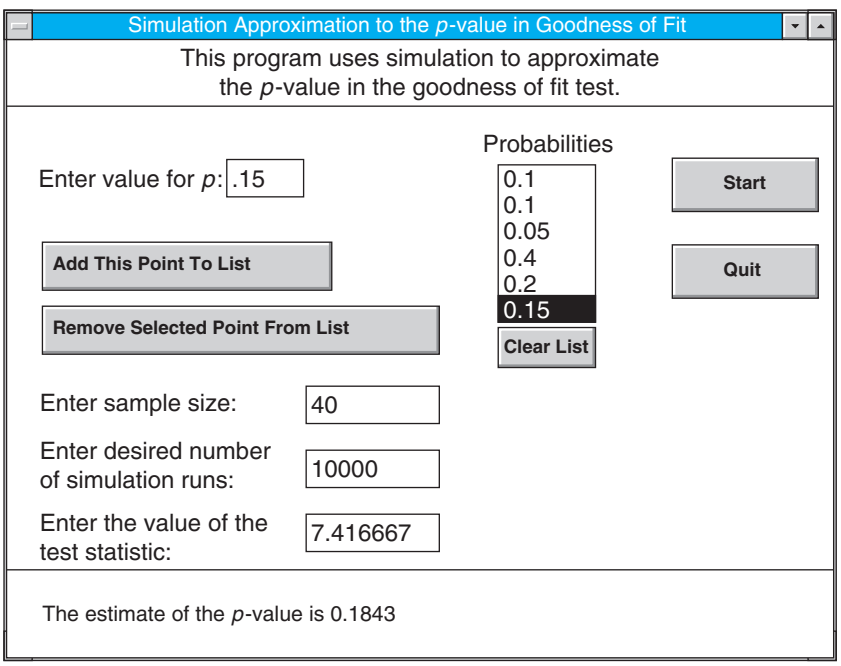

FIGURE 11.1

Since the number of the 10^4 simulated values that exceed 7.4167 is a binomial random variable with parameters $n = 10^4$ and $p = p$-value, it follows that a 90 percent confidence interval for the p-value is

$$p\text{-value} \in .1843 \pm 1.645\sqrt{.1843(.8157)/10^4}$$

That is, with 90 percent confidence

$$p\text{-value} \in (.1779, .1907) \quad \blacksquare$$

11.3 GOODNESS OF FIT TESTS WHEN SOME PARAMETERS ARE UNSPECIFIED

We can also perform goodness of fit tests of a null hypothesis that does not completely specify the probabilities $\{p_i, i = 1, \ldots, k\}$. For instance, consider the situation previously mentioned in which one is interested in testing whether the number of accidents occurring daily in a certain industrial plant is Poisson distributed with some unknown mean λ. To test this hypothesis, suppose that the daily number of accidents is recorded for n days — let $Y_1, \ldots, Y_n$ be these data. To analyze these data we must first address the difficulty that the Y_i can assume an infinite number of possible values. However, this is easily dealt with by breaking up the possible values into a finite number k of regions and then considering the region in which each Y_i falls. For instance, we might say that the outcome of the number of accidents on a given day is in region 1 if there are 0 accidents, region 2 if there is 1 accident, and region 3 if there are 2 or 3 accidents, region 4 if there are 4 or 5 accidents, and region 5 if there are more than 5 accidents. Hence, if the distribution is indeed Poisson with mean λ, then

$$\begin{aligned}
p_1 &= P\{Y = 0\} = e^{-\lambda} \qquad (11.3.1)\\
p_2 &= P\{Y = 1\} = \lambda e^{-\lambda}\\
p_3 &= P\{Y = 2\} + P\{Y = 3\} = \frac{e^{-\lambda}\lambda^2}{2} + \frac{e^{-\lambda}\lambda^3}{6}\\
p_4 &= P\{Y = 4\} + P\{Y = 5\} = \frac{e^{-\lambda}\lambda^4}{24} + \frac{e^{-\lambda}\lambda^5}{120}\\
p_5 &= P\{Y > 5\} = 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{e^{-\lambda}\lambda^2}{2} - \frac{e^{-\lambda}\lambda^3}{6} - \frac{e^{-\lambda}\lambda^4}{24} - \frac{e^{-\lambda}\lambda^5}{120}
\end{aligned}$$

The second difficulty we face in obtaining a goodness of fit test results from the fact that the mean value λ is not specified. Clearly, the intuitive thing to do is to assume that H_0 is true and then estimate it from the data — say, $\hat{\lambda}$ is the estimate of λ — and then

compute the test statistic

$$T = \sum_{i=1}^{k} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

where X_i is, as before, the number of Y_j that fall in region $i, i = 1, \ldots, k$, and $\hat{p}_i$ is the estimated probability of the event that Y_j falls in region i, which is determined by substituting $\hat{\lambda}$ for λ in expression 11.3.1 for p_i.

In general, this approach can be utilized whenever there are unspecified parameters in the null hypothesis that are needed to compute the quantities $p_i, i = 1, \ldots, k$. Suppose now that there are m such unspecified parameters and that they are to be estimated by the method of maximum likelihood. It can then be proven that when n is large, the test statistic T will have, when H_0 is true, approximately a chi-square distribution with $k - 1 - m$ degrees of freedom. (In other words, one degree of freedom is lost for each parameter that needs to be estimated.) The test is, therefore, to

$$\begin{array}{lll} \text{reject} & H_0 & \text{if} \quad T \geq \chi^2_{\alpha,k-1-m} \\ \text{accept} & H_0 & \text{otherwise} \end{array}$$

An equivalent way of performing the foregoing is to first determine the value of the test statistic T, say $T = t$, and then compute

$$p\text{-value} \approx P\{\chi^2_{k-1-m} \geq t\}$$

The hypothesis would be rejected if $\alpha \geq p$-value.

EXAMPLE 11.3a Suppose the weekly number of accidents over a 30-week period is as follows:

8	0	0	1	3	4	0	2	12	5
1	8	0	2	0	1	9	3	4	5
3	3	4	7	4	0	1	2	1	2

Test the hypothesis that the number of accidents in a week has a Poisson distribution.

SOLUTION Since the total number of accidents in the 30 weeks is 95, the maximum likelihood estimate of the mean of the Poisson distribution is

$$\hat{\lambda} = \frac{95}{30} = 3.16667$$

Since the estimate of $P\{Y = i\}$ is then

$$P\{Y = i\} \stackrel{est}{=} \frac{e^{-\hat{\lambda}}\hat{\lambda}^i}{i!}$$

we obtain, after some computation, that with the five regions as given in the beginning of this section,

$$\hat{p}_1 = .04214$$
$$\hat{p}_2 = .13346$$
$$\hat{p}_3 = .43434$$
$$\hat{p}_4 = .28841$$
$$\hat{p}_5 = .10164$$

Using the data values $X_1 = 6, X_2 = 5, X_3 = 8, X_4 = 6, X_5 = 5$, an additional computation yields the test statistic value

$$T = \sum_{i=1}^{5} \frac{(X_i - 30\hat{p}_i)^2}{30\hat{p}_i} = 21.99156$$

To determine the p-value, we run Program 5.8.1a. This yields

$$\begin{aligned} p\text{-value} &\approx P\{\chi_3^2 > 21.99\} \\ &= 1 - .999936 \\ &= .000064 \end{aligned}$$

and so the hypothesis of an underlying Poisson distribution is rejected. (Clearly, there were too many weeks having 0 accidents for the hypothesis that the underlying distribution is Poisson with mean 3.167 to be tenable.) ■

11.4 TESTS OF INDEPENDENCE IN CONTINGENCY TABLES

In this section, we consider problems in which each member of a population can be classified according to two distinct characteristics — which we shall denote as the X-characteristic and the Y-characteristic. We suppose that there are r possible values for the X-characteristic and s for the Y-characteristic, and let

$$P_{ij} = P\{X = i, Y = j\}$$

for $i = 1, \ldots, r, j = 1, \ldots, s$. That is, P_{ij} represents the probability that a randomly chosen member of the population will have X-characteristic i and Y-characteristic j.

The different members of the population will be assumed to be independent. Also, let

$$p_i = P\{X = i\} = \sum_{j=1}^{s} P_{ij}, \qquad i = 1, \ldots, r$$

and

$$q_j = P\{Y = j\} = \sum_{i=1}^{r} P_{ij}, \qquad j = 1, \ldots, s$$

That is, p_i is the probability that an arbitrary member of the population will have X-characteristic i, and q_j is the probability it will have Y-characteristic j.

We are interested in testing the hypothesis that a population member's X- and Y-characteristics are independent. That is, we are interested in testing

$$H_0 : P_{ij} = p_i q_j, \quad \text{for all} \quad \begin{matrix} i = 1, \ldots, r \\ j = 1, \ldots, s \end{matrix}$$

against the alternative

$$H_1 : P_{ij} \neq p_i q_j, \quad \text{for some} \quad \begin{matrix} i, j \quad i = 1, \ldots, r \\ j = 1, \ldots, s \end{matrix}$$

To test this hypothesis, suppose that n members of the population have been sampled, with the result that N_{ij} of them have simultaneously had X-characteristic i and Y-characteristic $j, i = 1, \ldots, r, j = 1, \ldots, s$.

Since the quantities $p_i, i = 1, \ldots, r$, and $q_j, j = 1, \ldots, s$ are not specified by the null hypothesis, they must first be estimated. Now since

$$N_i = \sum_{j=1}^{s} N_{ij}, \qquad i = 1, \ldots, r$$

represents the number of the sampled population members that have X-characteristic i, a natural (in fact, the maximum likelihood) estimator of p_i is

$$\hat{p}_i = \frac{N_i}{n}, \qquad i = 1, \ldots, r$$

Similarly, letting

$$M_j = \sum_{i=1}^{r} N_{ij}, \qquad j = 1, \ldots, s$$

denote the number of sampled members having Y-characteristic j, the estimator for q_j is

$$\hat{q}_j = \frac{M_j}{n}, \qquad j = 1, \ldots, s$$

At first glance, it may seem that we have had to use the data to estimate $r+s$ parameters. However, since the p_i's and q_j's have to sum to 1 — that is, $\sum_{i=1}^{r} p_i = \sum_{j=1}^{s} q_j = 1$ — we need estimate only $r-1$ of the p's and $s-1$ of the q's. (For instance, if r were equal to 2, then an estimate of p_1 would automatically provide an estimate of p_2 since $p_2 = 1 - p_1$.) Hence, we actually need estimate $r-1+s-1 = r+s-2$ parameters, and since each population member has $k = rs$ different possible values, it follows that the resulting test statistic will, for large n, have approximately a chi-square distribution with $rs - 1 - (r+s-2) = (r-1)(s-1)$ degrees of freedom.

Finally, since

$$\begin{aligned} E[N_{ij}] &= nP_{ij} \\ &= np_iq_j \qquad \text{when } H_0 \text{ is true} \end{aligned}$$

it follows that the test statistic is given by

$$T = \sum_{j=1}^{s}\sum_{i=1}^{r} \frac{(N_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j} = \sum_{j=1}^{s}\sum_{i=1}^{r} \frac{N_{ij}^2}{n\hat{p}_i\hat{q}_j} - n$$

and the approximate significance level α test is to

$$\begin{aligned} \text{reject} \quad & H_0 \quad \text{if} \quad T \geq \chi^2_{\alpha,(r-1)(s-1)} \\ \text{not reject} \quad & H_0 \quad \text{otherwise} \end{aligned}$$

EXAMPLE 11.4a A sample of 300 people was randomly chosen, and the sampled individuals were classified as to their gender and political affiliation, Democrat, Republican, or Independent. The following table, called a *contingency table*, displays the resulting data.

	j			
i	**Democrat**	**Republican**	**Independent**	**Total**
Women	68	56	32	156
Men	52	72	20	144
Total	120	128	52	300

Thus, for instance, the contingency table indicates that the sample of size 300 contained 68 women who classified themselves as Democrats, 56 women who classified themselves

as Republicans, and 32 women who classified themselves as Independents; that is, $N_{11} = 68$, $N_{12} = 56$, and $N_{13} = 32$. Similarly, $N_{21} = 52$, $N_{22} = 72$, and $N_{23} = 20$.

Use these data to test the hypothesis that a randomly chosen individual's gender and political affiliation are independent.

SOLUTION From the above data, we obtain that the six values of $n\hat{p}_i\hat{q}_j = N_iM_j/n$ are as follows:

$$\frac{N_1M_1}{n} = \frac{156 \times 120}{300} = 62.40$$
$$\frac{N_1M_2}{n} = \frac{156 \times 128}{300} = 66.56$$
$$\frac{N_1M_3}{n} = \frac{156 \times 52}{300} = 27.04$$
$$\frac{N_2M_1}{n} = \frac{144 \times 120}{300} = 57.60$$
$$\frac{N_2M_2}{n} = \frac{144 \times 128}{300} = 61.44$$
$$\frac{N_2M_3}{n} = \frac{144 \times 52}{300} = 24.96$$

The value of the test statistic is thus

$$\begin{aligned} TS &= \frac{(68 - 62.40)^2}{62.40} + \frac{(56 - 66.56)^2}{66.56} + \frac{(32 - 27.04)^2}{27.04} + \frac{(52 - 57.60)^2}{57.60} \\ &\quad + \frac{(72 - 61.44)^2}{61.44} + \frac{(20 - 24.96)^2}{24.96} \\ &= 6.433 \end{aligned}$$

Since $(r - 1)(s - 1) = 2$, we must compare the value of TS with the critical value $\chi^2_{.05,2}$. From Table A2

$$\chi^2_{.05,2} = 5.991$$

Since $TS \geq 5.991$, the null hypothesis is rejected at the 5 percent level of significance. That is, the hypothesis that gender and political affiliation of members of the population are independent is rejected at the 5 percent level of significance. ■

The results of the test of independence of the characteristics of a randomly chosen member of the population can also be obtained by computing the resulting p-value. If the observed value of the test statistic is $T = t$, then the significance level α test would call

for rejecting the hypothesis of independence if the p-value is less than or equal to α, where

$$\begin{aligned} p\text{-value} &= P_{H_0}\{T \geq t\} \\ &\approx P\{\chi^2_{(r-1)(s-1)} \geq t\} \end{aligned}$$

Program 11.4 will compute the value of T.

EXAMPLE 11.4b A company operates four machines on three separate shirts daily. The following contingency table presents the data during a 6-month time period, concerning the machine breakdowns that resulted.

Number of Breakdowns

	Machine				
	A	**B**	**C**	**D**	**Total per Shift**
Shift 1	10	12	6	7	35
Shift 2	10	24	9	10	53
Shift 3	13	20	7	10	50
Total per Machine	33	56	22	27	138

Suppose we are interested in determining whether a machine's breakdown probability during a particular shift is influenced by that shift. In other words, we are interested in testing, for an arbitrary breakdown, whether the machine causing the breakdown and the shift on which the breakdown occurred are independent.

SOLUTION A direct computation, or the use of Program 11.4, gives that the value of the test statistic is 1.8148 (see Figure 11.2). Utilizing Program 5.8.1a then gives that

$$\begin{aligned} p\text{-value} &\approx P\{\chi^2_6 \geq 1.8148\} \\ &= 1 - .0641 \\ &= .9359 \end{aligned}$$

and so the hypothesis that the machine that causes a breakdown is independent of the shift on which the breakdown occurs is accepted. ■

11.5 TESTS OF INDEPENDENCE IN CONTINGENCY TABLES HAVING FIXED MARGINAL TOTALS

In Example 11.4a, we were interested in determining whether gender and political affiliation were dependent in a particular population. To test this hypothesis, we first chose

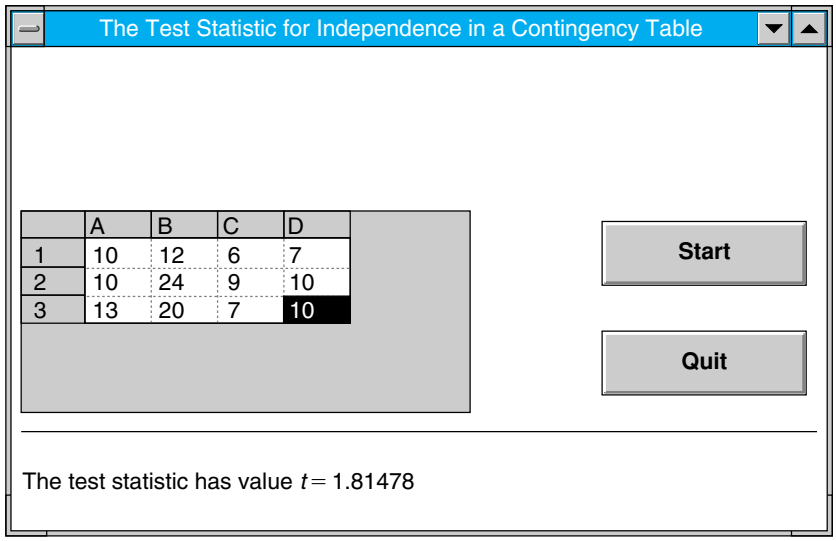

FIGURE 11.2

a random sample of people from this population and then noted their characteristics. However, another way in which we could gather data is to fix in advance the numbers of men and women in the sample and then choose random samples of those sizes from the subpopulations of men and women. That is, rather than let the numbers of women and men in the sample be determined by chance, we might decide these numbers in advance. Because doing so would result in fixed specified values for the total numbers of men and women in the sample, the resulting contingency table is often said to have *fixed margins* (since the totals are given in the margins of the table).

It turns out that even when the data are collected in the manner prescribed above, the same hypothesis test as given in Section 11.4 can still be used to test for the independence of the two characteristics. The test statistic remains

$$TS = \sum_i \sum_j \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

where

N_{ij} = number of members of sample who have both X-characteristic i and Y-characteristic j

N_i = number of members of sample who have X-characteristic i

M_j = number of members of sample who have Y-characteristic j

and

$$\hat{e}_{ij} = n\hat{p}_i\hat{q}_j = \frac{N_i M_j}{n}$$

where n is the total size of the sample.

In addition, it is still true that when H_0 is true, TS will approximately have a chi-square distribution with $(r-1)(s-1)$ degrees of freedom. (The quantities r and s refer, of course, to the numbers of possible values of the X- and Y-characteristic, respectively.) In other words, the test of the independence hypothesis is unaffected by whether the marginal totals of one characteristic are fixed in advance or result from a random sample of the entire population.

EXAMPLE 11.5a A randomly chosen group of 20,000 nonsmokers and one of 10,000 smokers were followed over a 10-year period. The following data relate the numbers of them that developed lung cancer during that period.

	Smokers	Nonsmokers	Total
Lung cancer	62	14	76
No lung cancer	9,938	19,986	29,924
Total	10,000	20,000	30,000

Test the hypothesis that smoking and lung cancer are independent. Use the 1 percent level of significance.

SOLUTION The estimates of the expected number to fall in each ij cell when smoking and lung cancer are independent are

$$\hat{e}_{11} = \frac{(76)(10{,}000)}{30{,}000} = 25.33$$

$$\hat{e}_{12} = \frac{(76)(20{,}000)}{30{,}000} = 50.67$$

$$\hat{e}_{21} = \frac{(29{,}924)(10{,}000)}{30{,}000} = 9{,}974.67$$

$$\hat{e}_{22} = \frac{(29{,}924)(20{,}000)}{30{,}000} = 19{,}949.33$$

Therefore, the value of the test statistic is

$$\begin{aligned} TS &= \frac{(62-25.33)^2}{25.33} + \frac{(14-50.67)^2}{50.67} + \frac{(9{,}938-9{,}974.67)^2}{9{,}974.67} \\ &\quad + \frac{(19{,}986-19{,}949.33)^2}{19{,}949.33} \\ &= 53.09 + 26.54 + .13 + .07 = 79.83 \end{aligned}$$

Since this is far larger than $\chi^2_{.01,1} = 6.635$, we reject the null hypothesis that whether a randomly chosen person develops lung cancer is independent of whether that person is a smoker. ■

We now show how to use the framework of this section to test the hypothesis that m discrete population distributions are equal. Consider m separate populations, each of whose members takes on one of the values $1, \ldots, n$. Suppose that a randomly chosen member of population i will have value j with probability

$$p_{i,j}, \qquad i = 1, \ldots, m, \qquad j = 1, \ldots, n$$

and consider a test of the null hypothesis

$$H_0 : p_{1,j} = p_{2,j} = p_{3,j} = \cdots = p_{m,j}, \qquad \text{for each } j = 1, \ldots, n$$

To obtain a test of this null hypothesis, consider first the superpopulation consisting of all members of each of the m populations. Any member of this superpopulation can be classified according to two characteristics. The first characteristic specifies which of the m populations the member is from, and the second characteristic specifies its value. The hypothesis that the population distributions are equal becomes the hypothesis that, for each value, the proportion of members of each population having that value are the same. But this is exactly the same as saying that the two characteristics of a randomly chosen member of the superpopulation are independent. (That is, the value of a randomly chosen superpopulation member is independent of the population to which this member belongs.)

Therefore, we can test H_0 by randomly choosing sample members from each population. If we let M_i denote the sample size from population i and let $N_{i,j}$ denote the number of values from that sample that are equal to j, $i = 1, \ldots, m$, $j = 1, \ldots, n$, then we can test H_0 by testing for independence in the following contingency table.

	Population				
Value	**1**	**2**	***i***	***m***	**Totals**
1	$N_{1,1}$	$N_{2,1}\ldots$	$N_{i,1}\ldots$	$N_{m,1}$	N_1
2					
⋮					
j	$N_{1,j}$	$N_{2,j}\ldots$	$N_{i,j}\ldots$	$N_{m,j}$	N_j
⋮					
n	$N_{1,n}$	$N_{2,n}\ldots$	$N_{i,n}$	$N_{m,n}$	N_n
Totals	M_1	$M_2\ldots$	$M_i\ldots$	M_m	

Note that N_j denotes the number of sampled members that have value j.

EXAMPLE 11.5b A recent study reported that 500 female office workers were randomly chosen and questioned in each of four different countries. One of the questions related to whether these women often received verbal or sexual abuse on the job. The following data resulted.

Country	Number Reporting Abuse
Australia	28
Germany	30
Japan	51
United States	55

Based on these data, is it plausible that the proportions of female office workers who often feel abused at work are the same for these countries?

SOLUTION Putting the above data in the form of a contingency table gives the following.

	Country				
	1	**2**	**3**	**4**	**Totals**
Receive abuse	28	30	58	55	171
Do not receive abuse	472	470	442	445	1,829
Totals	500	500	500	500	2,000

We can now test the null hypothesis by testing for independence in the preceding contingency table. If we run Program 11.4, then the value of the test statistic and the resulting p-value are

$$TS = 19.51, \qquad p\text{-value} \approx .0002$$

Therefore, the hypothesis that the percentages of women who feel they are being abused on the job are the same for these countries is rejected at the 1 percent level of significance (and, indeed, at any significance level above .02 percent). ■

*11.6 THE KOLMOGOROV–SMIRNOV GOODNESS OF FIT TEST FOR CONTINUOUS DATA

Suppose now that $Y_1, \ldots, Y_n$ represents sample data from a continuous distribution, and that we wish to test the null hypothesis H_0 that F is the population distribution, where F is a specified continuous distribution function. One approach to testing H_0 is to break up the set of possible values of the Y_j into k distinct intervals, say,

$$(y_0, y_1), (y_1, y_2), \ldots, (y_{k-1}, y_k), \qquad \text{where} \quad y_0 = -\infty, y_k = +\infty$$

and then consider the discretized random variables $Y_j^d, j = 1, \ldots, n$, defined by

$$Y_j^d = i \qquad \text{if } Y_j \text{ lies in the interval } (y_{i-1}, y_i)$$

The null hypothesis then implies that

$$P\{Y_j^d = i\} = F(y_i) - F(y_{i-1}), \qquad i = 1, \ldots, k$$

and this can be tested by the chi-square goodness of fit test already presented.

There is, however, another way of testing that the Y_j come from the continuous distribution function F that is generally more efficient than discretizing; it works as follows. After observing $Y_1, \ldots, Y_n$, let F_e be the empirical distribution function defined by

$$F_e(x) = \frac{\#i : Y_i \le x}{n}$$

That is, $F_e(x)$ is the proportion of the observed values that are less than or equal to x. Because $F_e(x)$ is a natural estimator of the probability that an observation is less than or equal to x, it follows that, if the null hypothesis that F is the underlying distribution is correct, it should be close to $F(x)$. Since this is so for all x, a natural quantity on which to base a test of H_0 is the test quantity

$$D \equiv \underset{x}{\text{Maximum}} |F_e(x) - F(x)|$$

where the maximum is over all values of x from $-\infty$ to $+\infty$. The quantity D is called the *Kolmogorov–Smirnov test statistic.*

---

* Optional section.

To compute the value of D for a given data set $Y_j = y_j, j = 1, \ldots, n$, let $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ denote the values of the y_j in increasing order. That is,

$$y_{(j)} = j\text{th smallest of } y_1, \ldots, y_n$$

For example, if $n = 3$ and $y_1 = 3, y_2 = 5, y_3 = 1$, then $y_{(1)} = 1, y_{(2)} = 3, y_{(3)} = 5$. Since $F_e(x)$ can be written

$$F_e(x) = \begin{cases} 0 & \text{if } x < y_{(1)} \\ \frac{1}{n} & \text{if } y_{(1)} \le x < y_{(2)} \\ \vdots & \\ \frac{j}{n} & \text{if } y_{(j)} \le x < y_{(j+1)} \\ \vdots & \\ 1 & \text{if } y_{(n)} \le x \end{cases}$$

we see that $F_e(x)$ is constant within the intervals $(y_{(j-1)}, y_{(j)})$ and then jumps by $1/n$ at the points $y_{(1)}, \ldots, y_{(n)}$. Since $F(x)$ is an increasing function of x that is bounded by 1, it follows that the maximum value of $F_e(x) - F(x)$ is nonnegative and occurs at one of the points $y_{(j)}, j = 1, \ldots, n$ (see Figure 11.3).

That is,

$$\underset{x}{\text{Maximum}}\{F_e(x) - F(x)\} = \underset{j=1,\ldots,n}{\text{Maximum}} \left\{ \frac{j}{n} - F(y_{(j)}) \right\} \tag{11.6.1}$$

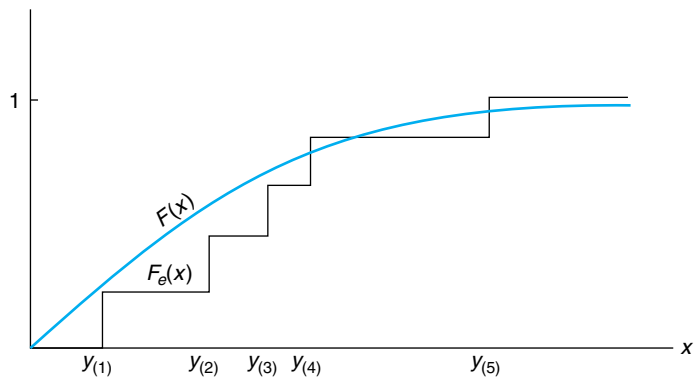

FIGURE 11.3 $n = 5$.

Similarly, the maximum value of $F(x) - F_e(x)$ is also nonnegative and occurs immediately before one of the jump points $y_{(j)}$; and so

$$\underset{x}{\text{Maximum}}\{F(x) - F_e(x)\} = \underset{j=1,\ldots,n}{\text{Maximum}} \left\{ F(y_{(j)}) - \frac{j-1}{n} \right\} \tag{11.6.2}$$

From Equations 11.6.1 and 11.6.2, we see that

$$\begin{aligned} D &= \underset{x}{\text{Maximum}} |F_e(x) - F(x)| \\ &= \text{Maximum}\{\text{Maximum}\{F_e(x) - F(x)\}, \text{Maximum}\{F(x) - F_e(x)\}\} \\ &= \text{Maximum} \left\{ \frac{j}{n} - F(y_{(j)}), F(y_{(j)}) - \frac{j-1}{n}, j = 1, \ldots, n \right\} \end{aligned} \tag{11.6.3}$$

Equation 11.6.3 can be used to compute the value of D.

Suppose now that the Y_j are observed and their values are such that $D = d$. Since a large value of D would appear to be inconsistent with the null hypothesis that F is the underlying distribution, it follows that the p-value for this data set is given by

$$p\text{-value} = P_F\{D \geq d\}$$

where we have written P_F to make explicit that this probability is to be computed under the assumption that H_0 is correct (and so F is the underlying distribution).

The above p-value can be approximated by a simulation that is made easier by the following proposition, which shows that $P_F\{D \geq d\}$ does not depend on the underlying distribution F. This result enables us to estimate the p-value by doing the simulation with any continuous distribution F we choose [thus allowing us to use the uniform (0, 1) distribution].

PROPOSITION 11.6.1

$P_F\{D \geq d\}$ is the same for any continuous distribution F.

Proof

$$\begin{aligned} P_F\{D \geq d\} &= P_F \left\{ \underset{x}{\text{Maximum}} \left| \frac{\#i : Y_i \leq x}{n} - F(x) \right| \geq d \right\} \\ &= P_F \left\{ \underset{x}{\text{Maximum}} \left| \frac{\#i : F(Y_i) \leq F(x)}{n} - F(x) \right| \geq d \right\} \\ &= P \left\{ \underset{x}{\text{Maximum}} \left| \frac{\#i : U_i \leq F(x)}{n} - F(x) \right| \geq d \right\} \end{aligned}$$

where $U_1, \ldots, U_n$ are independent uniform $(0, 1)$ random variables. The first equality following because F is an increasing function and so $Y \le x$ is equivalent to $F(Y) \le F(x)$; and the second because of the result (whose proof is left as an exercise) that if Y has the continuous distribution F then the random variable $F(Y)$ is uniform on (0, 1).

Continuing the above, we see by letting $y = F(x)$ and noting that as x ranges from $-\infty$ to $+\infty$, $F(x)$ ranges from 0 to 1, that

$$P_F\{D \ge d\} = P\left\{\underset{0\le y\le 1}{\text{Maximum}} \left|\frac{\#i : U_i \le y}{n} - y\right| \ge d\right\}$$

which shows that the distribution of D, when H_0 is true, does not depend on the actual distribution F. ■

It follows from the above proposition that after the value of D is determined from the data, say, $D = d$, the p-value can be obtained by doing a simulation with the uniform (0, 1) distribution. That is, we generate a set of n random numbers $U_1, \ldots, U_n$ and then check whether or not the inequality

$$\underset{0\le y\le 1}{\text{Maximum}} \left|\frac{\#i : U_i \le y}{n} - y\right| \ge d$$

is valid. This is then repeated many times and the proportion of times that it is valid is our estimate of the p-value of the data set. As noted earlier, the left side of the inequality can be computed by ordering the random numbers and then using the identity

$$\text{Max}\left|\frac{\#i : U_i \le y}{n} - y\right| = \text{Max}\left\{\frac{j}{n} - U_{(j)}, U_{(j)} - \frac{(j-1)}{n}, j = 1, \ldots, n\right\}$$

where $U_{(j)}$ is the jth smallest value of $U_1, \ldots, U_n$. For example, if $n = 3$ and $U_1 = .7$, $U_2 = .6$, $U_3 = .4$, then $U_{(1)} = .4$, $U_{(2)} = .6$, $U_{(3)} = .7$ and the value of D for this data set is

$$D = \text{Max}\left\{\frac{1}{3} - .4,\ \frac{2}{3} - .6,\ 1 - .7,\ .4,\ .6 - \frac{1}{3},\ .7 - \frac{2}{3}\right\} = .4$$

A significance level α test can be obtained by considering the quantity D^* defined by

$$D^* = (\sqrt{n} + .12 + .11/\sqrt{n})D$$

Letting d_α^* be such that

$$P_F\{D^* \ge d_\alpha^*\} = \alpha$$

then the following are accurate approximations for d_α^* for a variety of values:

$$d_{.1}^* = 1.224, \qquad d_{.05}^* = 1.358, \qquad d_{.025}^* = 1.480, \qquad d_{.01}^* = 1.626$$

The level α test would reject the null hypothesis that F is the distribution if the observed value of D^* is at least as large as d_α^*.

EXAMPLE 11.6a Suppose we want to test the hypothesis that a given population distribution is exponential with mean 100; that is, $F(x) = 1 - e^{-x/100}$. If the (ordered) values from a sample of size 10 from this distribution are

$$66, 72, 81, 94, 112, 116, 124, 140, 145, 155$$

what conclusion can be drawn?

SOLUTION To answer the above, we first employ Equation 11.6.3 to compute the value of the Kolmogorov–Smirnov test quantity D. After some computation this gives the result $D = .4831487$, which results in

$$D^* = .48315(\sqrt{10} + 0.12 + 0.11/\sqrt{10}) = 1.603$$

Because this exceeds $d_{.025}^* = 1.480$, it follows that the null hypothesis that the data come from an exponential distribution with mean 100 would be rejected at the 2.5 percent level of significance. (On the other hand, it would not be rejected at the 1 percent level of significance.) ■

Problems

1. According to the Mendelian theory of genetics, a certain garden pea plant should produce either white, pink, or red flowers, with respective probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$. To test this theory, a sample of 564 peas was studied with the result that 141 produced white, 291 produced pink, and 132 produced red flowers. Using the chi-square approximation, what conclusion would be drawn at the 5 percent level of significance?

2. To ascertain whether a certain die was fair, 1,000 rolls of the die were recorded, with the following results.

Outcome	Number of Occurrences
1	158
2	172
3	164
4	181
5	160
6	165

Test the hypothesis that the die is fair (that is, that $p_i = \frac{1}{6}, i = 1, \ldots, 6$) at the 5 percent level of significance. Use the chi-square approximation.

3. Determine the birth and death dates of 100 famous individuals and, using the four-category approach of Example 11.2a, test the hypothesis that the death month is not affected by the birth month. Use the chi-square approximation.

4. It is believed that the daily number of electrical power failures in a certain Midwestern city is a Poisson random variable with mean 4.2. Test this hypothesis if over 150 days the number of days having i power failures is as follows:

Failures	Number of Days
0	0
1	5
2	22
3	23
4	32
5	22
6	19
7	13
8	6
9	4
10	4
11	0

5. Among 100 vacuum tubes tested, 41 had lifetimes of less than 30 hours, 31 had lifetimes between 30 and 60 hours, 13 had lifetimes between 60 and 90 hours, and 15 had lifetimes of greater than 90 hours. Are these data consistent with the hypothesis that a vacuum tube's lifetime is exponentially distributed with a mean of 50 hours?

6. The past output of a machine indicates that each unit it produces will be

top grade	with probability	.40
high grade	with probability	.30
medium grade	with probability	.20
low grade	with probability	.10

A new machine, designed to perform the same job, has produced 500 units with the following results.

top grade	234
high grade	117
medium grade	81
low grade	68

Can the difference in output be ascribed solely to chance?

7. The neutrino radiation from outer space was observed during several days. The frequencies of signals were recorded for each sidereal hour and are as given below:

Frequency of Neutrino Radiation from Outer Space

Hour Starting at	Frequency of Signals	Hour Starting at	Frequency of Signals
0	24	12	29
1	24	13	26
2	36	14	38
3	32	15	26
4	33	16	37
5	36	17	28
6	41	18	43
7	24	19	30
8	37	20	40
9	37	21	22
10	49	22	30
11	51	23	42

Test whether the signals are uniformly distributed over the 24-hour period.

8. Neutrino radiation was observed over a certain period and the number of hours in which 0, 1, 2,. . . signals were received was recorded.

Number of Signals per Hour	Number of Hours with This Frequency of Signals
0	1,924
1	541
2	103
3	17
4	1
5	1
6 or more	0

Test the hypothesis that the observations come from a population having a Poisson distribution with mean .3.

9. In a certain region, insurance data indicate that 82 percent of drivers have no accidents in a year, 15 percent have exactly 1 accident, and 3 percent have 2 or more accidents. In a random sample of 440 engineers, 366 had no accidents, 68 had exactly 1 accident, and 6 had 2 or more. Can you conclude that engineers follow an accident profile that is different from the rest of the drivers in the region?

10. A study was instigated to see if southern California earthquakes of at least moderate size (having values of at least 4.4 on the Richter scale) are more likely to occur on certain days of the week than on others. The catalogs yielded the following data on 1,100 earthquakes.

Day	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
Number of Earthquakes	156	144	170	158	172	148	152

Test, at the 5 percent level, the hypothesis that an earthquake is equally likely to occur on any of the 7 days of the week.

11. Sometimes reported data fit a model so well that it makes one suspicious that the data are not being accurately reported. For instance, a friend of mine has reported that he tossed a fair coin 40,000 times and obtained 20,004 heads and 19,996 tails. Is such a result believable? Explain your reasoning.

12. Use simulation to determine the p-value and compare it with the result you obtained using the chi-square approximation in Problem 1. Let the number of simulation runs be

(a) 1,000;
(b) 5,000;
(c) 10,000.

13. A sample of size 120 had a sample mean of 100 and a sample standard deviation of 15. Of these 120 data values, 3 were less than 70; 18 were between 70 and 85; 30 were between 85 and 100; 35 were between 100 and 115; 32 were between 115 and 130; and 2 were greater than 130. Test the hypothesis that the sample distribution was normal.

14. In Problem 4, test the hypothesis that the daily number of failures has a Poisson distribution.

15. A random sample of 500 migrant families was classified by region and income (in units of $1,000). The following data resulted.

Income	South	North
0–10	42	53
10–20	55	90
20–30	47	88
>30	36	89

Determine the p-value of the test that a family's income and region are independent.

16. The following data relate the mother's age and the birthweight (in grams) of her child.

	Birthweight	
Maternal Age	**Less Than 2,500 Grams**	**More Than 2,500 Grams**
20 years or less	10	40
Greater than 20	15	135

Test the hypothesis that the baby's birthweight is independent of the mother's age.

17. Repeat Problem 16 with all of the data values doubled — that is, with these data:

20	80
30	270

18. The number of infant mortalities as a function of the baby's birthweight (in grams) for 72,730 live white births in New York in 1974 is as follows:

	Outcome at the End of 1 Year	
Birthweight	**Alive**	**Dead**
Less than 2,500	4,597	618
Greater than 2,500	67,093	422

Test the hypothesis that the birthweight is independent of whether or not the baby survives its first year.

19. An experiment designed to study the relationship between hypertension and cigarette smoking yielded the following data.

	Nonsmoker	**Moderate Smoker**	**Heavy Smoker**
Hypertension	20	38	28
No hypertension	50	27	18

Test the hypothesis that whether or not an individual has hypertension is independent of how much that person smokes.

20. The following table shows the number of defective, acceptable, and superior items in samples taken both before and after the introduction of a modification in the manufacturing process.

Defective	Acceptable	Superior
Before 25	218	22
After 9	103	14

Is this change significant at the .05 level?

21. A sample of 300 cars having cellular phones and one of 400 cars without phones were tracked for 1 year. The following table gives the number of these cars involved in accidents over that year.

	Accident	No Accident
Cellular phone	22	278
No phone	26	374

Use the above to test the hypothesis that having a cellular phone in your car and being involved in an accident are independent. Use the 5 percent level of significance.

22. To study the effect of fluoridated water supplies on tooth decay, two communities of roughly the same socioeconomic status were chosen. One of these communities had fluoridated water while the other did not. Random samples of 200 teenagers from both communities were chosen, and the numbers of cavities they had were determined. The following data resulted.

Cavities	Fluoridated Town	Nonfluoridated Town
0	154	133
1	20	18
2	14	21
3 or more	12	28

Do these data establish, at the 5 percent level of significance, that the number of dental cavities a person has is not independent of whether that person's water supply is fluoridated? What about at the 1 percent level?

23. To determine if a malpractice lawsuit is more likely to follow certain types of surgeries, random samples of three different types of surgeries were studied, and the following data resulted.

Type of Operation	Number Sampled	Number Leading to a Lawsuit
Heart surgery	400	16
Brain surgery	300	19
Appendectomy	300	7

Test the hypothesis that the percentages of the surgical operations that lead to lawsuits are the same for each of the three types.

(a) Use the 5 percent level of significance.
(b) Use the 1 percent level of significance.

24. In a famous article (S. Russell, "A red sky at night. . . ," *Metropolitan Magazine London*, **61**, p. 15, 1926) the following data set of frequencies of sunset colors and whether each was followed by rain was presented.

Sky Color	Number of Observations	Number Followed by Rain
Red	61	26
Mainly red	194	52
Yellow	159	81
Mainly yellow	188	86
Red and yellow	194	52
Gray	302	167

Test the hypothesis that whether it rains tomorrow is independent of the color of today's sunset.

25. Data are said to be from a *lognormal* distribution with parameters μ and σ if the natural logarithms of the data are normally distributed with mean μ and standard deviation σ. Use the Kolmogorov–Smirnov test with significance level .05 to decide whether the following lifetimes (in days) of a sample of cancer-bearing mice that have been treated with a certain cancer therapy might come from a lognormal distribution with parameters $\mu = 3$ and $\sigma = 4$.

$$24, 12, 36, 40, 16, 10, 12, 30, 38, 14, 22, 18$$