12.5: CHI-SQUARE GOODNESS OF FIT TESTS

In this section, the χ^2 distribution is used for testing the goodness of fit of a set of data to a specific probability distribution. In a test of goodness of fit, the actual frequencies in a category are compared to the frequencies that theoretically would be expected to occur if the data followed the specific probability distribution of interest.

Several steps are required to carry out a chi-square goodness of fit test. First, determine the specific probability distribution to be fitted to the data. Second, hypothesize or estimate the values of each parameter of the selected probability distribution (such as the mean). Next, determine the theoretical probability in each category using the selected probability distribution. Finally, use the χ^2 test statistic, Equation (12.8), to test whether the selected distribution is a good fit to the data.

The Chi-Square Goodness of Fit Test for a Poisson Distribution

Recall from Section 5.5 that the Poisson distribution was used to model the number of arrivals per minute at a bank located in the central business district of a city. Suppose that the actual arrivals per minute were observed in 200 one-minute periods over the course of a week. The results are summarized in Table 12.12.

TABLE 12.12
Frequency distribution of arrivals per minute during a lunch period

ARRIVALS	FREQUENCY
0	14
1	31
2	47
3	41
4	29
5	21
6	10
7	5
8	2
	200

To determine whether the number of arrivals per minute follows a Poisson distribution, the null and alternative hypotheses are as follows:

H_0: The number of arrivals per minute follows a Poisson distribution

H_1: The number of arrivals per minute does not follow a Poisson distribution

Since the Poisson distribution has one parameter, its mean λ, either a specified value can be included as part of the null and alternative hypotheses, or the parameter can be estimated from the sample data.

In this example, to estimate the average number of arrivals, you need to refer back to Equation (3.15) on page 111. Using Equation (3.15) and the computations in Table 12.13,

$$\bar{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}$$

$$\bar{X} = \frac{580}{200} = 2.90$$

TABLE 12.13
Computation of the sample average number of arrivals from the frequency distribution of arrivals per minute

ARRIVALS	FREQUENCY f_j	$m_j f_j$
0	14	0
1	31	31
2	47	94
3	41	123
4	29	116
5	21	105
6	10	60
7	5	35
8	2	16
	200	580

This value of the sample mean is used as the estimate of λ for the purposes of finding the probabilities from the tables of the Poisson distribution (Table E.7). From Table E.7, for $\lambda = 2.9$, the frequency of X successes ($X = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9$, or more) can be determined. The theoretical frequency for each is obtained by multiplying the appropriate Poisson probability by the sample size n. These results are summarized in Table 12.14.

TABLE 12.14
Actual and theoretical frequencies of the arrivals per minute

ARRIVALS	ACTUAL FREQUENCY f_0	PROBABILITY, $P(X)$, FOR POISSON DISTRIBUTION WITH $\lambda = 2.9$	THEORETICAL FREQUENCY $f_e = n \cdot P(X)$
0	14	0.0550	11.00
1	31	0.1596	31.92
2	47	0.2314	46.28
3	41	0.2237	44.74
4	29	0.1622	32.44
5	21	0.0940	18.80
6	10	0.0455	9.10
7	5	0.0188	3.76
8	2	0.0068	1.36
9 or more	0	0.0030	0.60

Observe from Table 12.14 that the theoretical frequency of 9 or more arrivals is less than 1.0. In order to have all categories contain a frequency of 1.0 or greater, the category 9 or more is combined with the category of 8 arrivals.

The chi-square test for determining whether the data follow a specific probability distribution is computed using Equation (12.8).

$$\chi^2_{k-p-1} = \sum_k \frac{(f_0 - f_e)^2}{f_e} \qquad \textbf{(12.8)}$$

where

f_0 = observed frequency
f_e = theoretical or expected frequency
k = number of categories or classes remaining after combining classes
p = number of parameters estimated from the data

Returning to the example concerning the arrivals at the bank, nine categories remain (0, 1, 2, 3, 4, 5, 6, 7, 8 or more). Since the mean of the Poisson distribution has been estimated from the data, the number of degrees of freedom are

$$k - p - 1 = 9 - 1 - 1 = 7 \text{ degrees of freedom}$$

Using the 0.05 level of significance, from Table E.4, the critical value of χ^2 with 7 degrees of freedom is 14.067. The decision rule is

$$\text{Reject } H_0 \text{ if } \chi^2 > 14.067\text{; otherwise do not reject } H_0.$$

From Table 12.15, since $\chi^2 = 2.28954 < 14.067$, the decision is not to reject H_0. There is insufficient evidence to conclude that the arrivals per minute do not fit a Poisson distribution.

TABLE 12.15
Computation of the chi-square test statistic for the arrivals per minute

ARRIVALS	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
0	14	11.00	3.00	9.0000	0.81818
1	31	31.92	−0.92	0.8464	0.02652
2	47	46.28	0.72	0.5184	0.01120
3	41	44.74	−3.74	13.9876	0.31264
4	29	32.44	−3.44	11.8336	0.36478
5	21	18.80	2.20	4.8400	0.25745
6	10	9.10	0.90	0.8100	0.08901
7	5	3.76	1.24	1.5376	0.40894
8 or more	2	1.96	0.04	0.0016	0.00082
					2.28954

The Chi-Square Goodness of Fit Test for a Normal Distribution

In Chapters 8 through 11, when testing hypotheses about numerical variables, the assumption is made that the underlying population was normally distributed. While graphical tools such as the box-and-whisker plot and the normal probability plot can be used to evaluate the validity of this assumption, an alternative that can be used with large sample sizes is the chi-square goodness-of-fit test for a normal distribution.

As an example of how the chi-square goodness-of-fit test for a normal distribution can be used, return to the 5-year annualized return rates achieved by the 158 growth funds summarized in Table 2.2 on page 45. Suppose you would like to test whether these returns follow a normal distribution. The null and alternative hypotheses are as follows:

H_0: The 5-year annualized return rates follow a normal distribution

H_1: The 5-year annualized return rates do not follow a normal distribution

In the case of the normal distribution, there are two parameters, the mean μ and the standard deviation σ, that can be estimated from the sample. For these data, $\overline{X} = 10.149$ and $S = 4.773$. Table 2.2 uses class interval widths of 5 with class boundaries beginning at −10.0. Since the normal distribution is continuous, the area in each class interval must be determined. In addition, since a normally distributed variable theoretically ranges from −∞ to +∞, the area beyond the class interval must also be accounted for. Thus, the area below −10 is the area below the Z value

$$Z = \frac{-10.0 - 10.149}{4.773} = -4.22$$

From Table E.2, the area below $Z = -4.22$ is approximately 0.0000.

To compute the area between −10.0 and −5.0, the area below −5.0 is computed as follows

$$Z = \frac{-5.0 - 10.149}{4.773} = -3.17$$

From Table E.2, the area below $Z = -3.17$ is approximately 0.00076. Thus, the area between −5.0 and −10.0 is the difference in the area below −5.0 and the area below −10.0, which is 0.00076 − 0.0000 = 0.00076.

Continuing, to compute the area between −5.0 and 0.0, the area below 0.0 is computed as follows

$$Z = \frac{0.0 - 10.149}{4.773} = -2.13$$

From Table E.2, the area below $Z = -2.13$ is approximately 0.0166. Thus the area between 0.0 and −5.0 is the difference in the area below 0.0 and the area below −5.0, which is 0.0166 − 0.00076 = 0.01584.

In a similar manner, the area in each class interval can be computed. The complete set of computations needed to find the area and expected frequency in each class is summarized in Table 12.16.

TABLE 12.16
Computation of the area and expected frequencies in each class interval for the 5-year annualized returns

CLASSES	X	$X - \bar{X}$	Z	AREA BELOW	AREA IN CLASS	$f_e = n \cdot P(X)$
Below −10.0	−10.0	−20.149	−4.22	0.00000	0.00000	0.00000
−10.0 but <−5.0	−5.0	−15.149	−3.17	0.00076	0.00076	0.12008
−5.0 but <0.0	0.0	−10.149	−2.13	0.01660	0.01584	2.50272
0.0 but <5.0	5.0	−5.149	−1.08	0.14010	0.12350	19.51300
5.0 but <10.0	10.0	−0.149	−0.03	0.48800	0.34790	54.96820
10.0 but <15.0	15.0	4.851	1.02	0.84610	0.3581	56.5798
15.0 but <20.0	20.0	9.851	2.06	0.98030	0.13420	21.20360
20.0 but <25.0	25.0	14.851	3.11	0.99906	0.01876	2.96408
25.0 but <30.0	30.0	19.851	4.16	1.00000	0.00094	0.14852
30.0 or more	—	—	+∞	1.00000	0.00000	0.00000

Observe from Table 12.16 that the theoretical frequency of below −10.0, between −10.0 and −5.0, between 25.0 and 30.0, and 30.0 or more are all less than 1.0. In order to have all categories contain a frequency of 1.0 or greater, the categories below −10.0 and between −10.0 and −5.0 are combined with the category −5.0 to 0.0 and the categories between 25.0 and 30.0, and 30.0 or more are combined with the category 20.0 to 25.0.

The chi-square test for determining whether the data follow a specific probability distribution is computed using Equation (12.8) on page CD12-2. In this example, after combining classes, 6 classes remain. Since the population mean and standard deviation have been estimated from the sample data, the number of degrees of freedom is equal to $k - p - 1 = 6 - 2 - 1 = 3$. Using a level of significance of 0.05, the critical value of chi-square with 3 degrees of freedom is 7.815. Table 12.17 summarizes the computations for the chi-square test.

TABLE 12.17
Computation of the chi-square test statistic for the 5-year annualized returns

CLASSES	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Below <0.0	4	2.6228	1.3772	1.89668	0.72315
0.0 but <5.0	14	19.5130	5.5130	30.39317	1.55759
5.0 but <10.0	58	54.9682	3.0318	9.19181	0.16722
10.0 but <15.0	61	56.5798	4.4202	19.53817	0.34532
15.0 but <20.0	17	21.2036	−4.2036	17.67025	0.83336
20.0 and above	4	3.1126	0.8874	0.78748	0.25300
					3.87963

From Table 12.17, since $\chi^2 = 3.87963 < 7.815$, the decision is not to reject H_0. Thus there is insufficient evidence to conclude that the 5-year annualized return does not fit a normal distribution.

PROBLEMS FOR SECTION 12.5

12.49 The manager of a computer network has collected data on the number of times that service has been interrupted on each day over the past 500 days. The results are as follows:

INTERRUPTIONS PER DAY	NUMBER OF DAYS
0	160
1	175
2	86
3	41
4	18
5	12
6	8
	500

Does the distribution of service interruptions follow a Poisson distribution? (Use the 0.01 level of significance.)

12.50 Referring to the data in problem 12.49, at the 0.01 level of significance, does the distribution of service interruptions follow a Poisson distribution with a population mean of 1.5 interruptions per day?

12.51 The manager of a commercial mortgage department of a large bank has collected data during the past two years concerning the number of commercial mortgages approved per week. The results from these two years (104 weeks) indicated the following:

NUMBER OF COMMERCIAL MORTGAGES APPROVED	FREQUENCY
0	13
1	25
2	32
3	17
4	9
5	6
6	1
7	1
	104

Does the distribution of commercial mortgages approved per week follow a Poisson distribution? (Use the 0.01 level of significance.)

12.52 A random sample of 500 car batteries revealed the following distribution of battery life (in years).

LIFE (IN YEARS)	FREQUENCY
0–under 1	12
1–under 2	94
2–under 3	170
3–under 4	188
4–under 5	28
5–under 6	8
	500

For these data, $\overline{X} = 2.80$ and $S = 0.97$. At the 0.05 level of significance, does battery life follow a normal distribution?

12.53 A random sample of 500 long distance telephone calls revealed the following distribution of call length (in minutes).

LENGTH (IN MINUTES)	FREQUENCY
0–under 5	48
5–under 10	84
10–under 15	164
15–under 20	126
20–under 25	50
25–under 30	28
	500

a. Compute the mean and standard deviation of this frequency distribution.
b. At the 0.05 level of significance, does call length follow a normal distribution?