

**Part One**  
**Collecting and Describing Data**

# 11

## The Chi-square Test

### 11.1 The Chi-squared ( $\chi^2$ ) Test

Consider a population that can be classified into several categories with respect to two attributes say, age and job performance. We may wish to test whether any or all of the categories named have any association with the two attributes.

Many times, managers need to know whether the differences they observe among several sample proportions are significant or only due to chance. Suppose a campaign manager for a particular candidate at a local government elections studies three geographically different regions and finds that 35 percent, 42 percent, and 51 percent of those voters surveyed in the three regions, respectively, recognize the candidate's name. If this difference is significant, the manager may conclude that location will affect the way the candidate should act. If the difference is not significant (that is, if the campaign manager concludes that the difference is solely due to chance), then he may decide that the place chosen to make a particular policy-making speech will have no effect on its reception. To run the campaign successfully, the manager needs to determine whether location and acceptance are dependent or independent.

The statistical technique used to carry out this test is called the *chi-squared test*. This is pronounced, "kya-square" and is denoted by the symbol  $\chi^2$ .

### 11.2 Contingency Tables

Suppose that in four regions, the Ministry of Health samples its hospital employees' attitudes toward job performance reviews. Respondents are given a choice between the present method (two reviews a year) and a proposed new method (quarterly reviews).

The following table shows the sample responses.

Employees	Regions				Total
	1	2	3	4	
Number who prefer present method	68	75	57	79	279
Number who prefer new method	32	45	33	31	141
<b>Total</b>	100	120	90	110	420

This table is called a *contingency table*. A contingency table is made up of rows and columns. The rows provide one basis of classification – preference for review methods, and the columns provide another basis of classification – geographical regions. This is called a “2 × 4 contingency table,” because it consists of two rows and four columns. It is usual to describe the dimensions of a contingency table by first stating the number of rows and then the number of columns. The “Total” column and the “Total” row are not counted as part of the dimensions.

### 11.3 The Chi-squared Statistic

It is required to go beyond the intuitive feelings about the observed and expected results of situations as described above.

The chi-squared statistic states that

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

where  $O_i$  is the  $i^{\text{th}}$  observed frequency

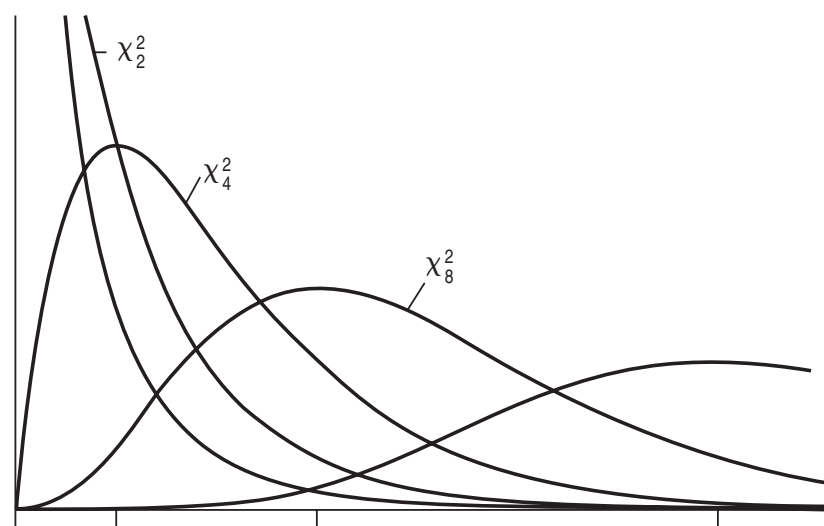
$E_i$  is the  $i^{\text{th}}$  expected or theoretical frequency.

The value of the chi-squared statistic can never be negative since the differences between the observed and expected frequencies are squared.

### 11.4 The Chi-squared Distribution

For the chi-squared test it is usual to take the null hypothesis that there is no association between the categories or classes represented by a contingency table. If the null hypothesis is true, then the sampling distribution of the chi-squared statistic,  $\chi^2$ , can be closely approximated by a continuous curve known as a *chi-squared distribution*. As in the case of the  $t$ -distribution, there is a different chi-squared distribution for each different number of degrees of freedom. Figure indicates three different chi-squared distributions that would correspond to 2, 4 and 8 degrees of freedom. For very small numbers of degrees of freedom, the

chi-squared distribution is severely skewed to the right. As the number of degrees of freedom increases, the curve rapidly becomes more symmetrical until the number reaches large values, at which point the distribution can be approximated by the normal.



The chi-squared distribution is a probability distribution. Therefore, the total area under the curve in each chi-squared distribution is 1.0. Like the  $t$ -distribution, so many chi-squared distributions are possible that it is not practical to construct a table that illustrates the areas under the curve for all possible values of the area.

The following table is an extract from a chi-squared distribution showing the appropriate chi-squared values for respective chi-squared statistics and degrees of freedom.

The values in the table are those, which a random variable with the  $\chi^2$  distribution on  $\nu$  degrees of freedom exceeds with the probability shown.

$\nu$	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.705	3.841	5.024	6.635	7.789
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589

## 11.6 Using the Chi-squared Test

It is first required to calculate the expected or theoretical frequency for each corresponding observed frequency. The expected frequency for  $C_{ij}$  is found by the formula

$$\frac{\text{Row}_i \text{ total} \times \text{Column}_j \text{ total}}{\text{Grand total}}$$

where  $C_{ij}$  is the cell for the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column.

### Example 1

A teacher is trying to pinpoint whether newspaper readership in a community is related to readers' educational achievement. A survey questioned adults in the area on their level of education and their frequency of readership. The results are shown in the following table.

Frequency of Readership	Level of Educational Achievement			
	Professional or Postgraduate	Undergraduate	High School Graduate	Did not finish High School
Never	7	14	13	16
Sometimes	13	17	7	7
Daily or Sunday editions only	39	41	10	5
Both editions	22	23	8	12

At the 0.05 significance level, does the frequency of newspaper readership in the community differ according to the readers' level of education?

### Solution 1

$H_0$ : There is no association between frequency of readership and the level of educational achievement.

$H_1$ : The frequency of readership and the level of educational achievement are not independent.

At 0.05 significant level, from the chi-squared tables, reject  $H_0$  if the chi-square statistic with  $(4 - 1)(4 - 1)$  degrees of freedom is more than 16.919.

The following table shows the calculation for the chi-squared statistic.

Row	Column	$\frac{R_i \times C_j}{G}$	$(O_i - E_i)$	$\frac{(O_i - E_i)^2}{E_i}$
1	1	$\frac{50 \times 81}{254}$	7 - 15.9	4.9818
1	2	$\frac{50 \times 95}{254}$	14 - 18.7	1.1813
1	3	$\frac{50 \times 38}{254}$	13 - 7.5	4.0333
1	4	$\frac{50 \times 40}{254}$	16 - 7.9	8.3051
2	1	$\frac{44 \times 81}{254}$	13 - 14.0	0.0714
2	2	$\frac{44 \times 95}{254}$	17 - 16.5	0.0152
2	3	$\frac{44 \times 38}{254}$	7 - 6.6	0.0242
2	4	$\frac{44 \times 40}{254}$	7 - 6.9	0.0014
3	1	$\frac{95 \times 81}{254}$	39 - 30.3	2.4980
3	2	$\frac{95 \times 95}{254}$	41 - 35.5	0.8521
3	3	$\frac{95 \times 38}{254}$	10 - 14.2	1.2423
3	4	$\frac{95 \times 40}{254}$	5 - 15	6.6667
4	1	$\frac{65 \times 81}{254}$	22 - 20.7	0.0816
4	2	$\frac{65 \times 95}{254}$	23 - 24.3	0.0695
4	3	$\frac{65 \times 38}{254}$	8 - 9.7	0.2979
4	4	$\frac{65 \times 40}{254}$	12 - 10.2	0.3176
				$\sum \frac{(O_i - E_i)^2}{E_i} = 30.639$

## 248 The Chi-square Test

Since  $\chi^2_{\text{statistic}}$  is greater than  $\chi^2_{(0.05)(9)}$  the  $H_0$  is not accepted. There is evidence that the frequency of readership is not independent of the level of educational achievement.

To use a chi-squared hypothesis test, we must have a sample size large enough to guarantee the similarity between the expected (theoretically correct) distribution and our sampling distribution of the  $\chi^2$ , the chi-squared statistic. When the expected frequencies are too small, the value of  $\chi^2$  will be overestimated and will result in too many rejections of the null hypothesis. To avoid making incorrect inferences from  $\chi^2$  hypothesis tests, follow the general rule that an expected frequency of less than 5 in one cell of a contingency table is too small to use.

### Example 2

A pharmaceutical company is testing four different tranquilizing drugs for their effects on driving skill. Subjects take a simulated driving test, and their scores reflect their errors. The more severe errors lead to higher scores. The results of these tests produced the following table:

Drug	Scores (errors)				Total
	A	B	C	D	
Drug 1	245	258	239	241	983
Drug 2	277	276	263	274	1090
Drug 3	215	232	225	247	919
Drug 4	241	253	237	246	977
<b>Total</b>	978	1019	964	1008	3969

At the 0.05 level of significance, do the four drugs affect driving skill differently?

### Solution 2

Drug	Expected Scores (errors)				Total
	A	B	C	D	
Drug 1	242.2	252.4	238.8	249.6	983
Drug 2	268.6	279.8	264.7	276.9	1090
Drug 3	226.5	235.9	223.2	233.4	919
Drug 4	240.7	250.9	237.3	248.1	977
<b>Total</b>	978	1019	964	1008	3969

**Example 4**

A certain company requires that undergraduates who are seeking positions with it to be interviewed by three different executives. This enables the company to obtain a consensus evaluation of each candidate. Each executive gives the candidate either a positive or a negative rating. The following table shows the results of the last 100 candidates:

<b>Possible Positive Ratings (3 interviews)</b>	0	1	2	3
<b>Number of candidates</b>	18	47	24	11

For manpower planning purposes, the director of recruitment for this company thinks that the interview process can be approximated by a binomial distribution with  $p = 0.40$ , that is, with a 40 percent chance of any candidate receiving a positive rating on any of one interview. If the director wants to test this hypothesis at the 0.10 level of significance, how should he proceed?

**Solution 4**

To solve this problem, we must determine whether the discrepancies between the observed frequencies and those we would expect (if the binomial distribution is the proper model to use) are actually due to chance. We can begin by determining what the binomial probabilities would be for this interview situation. For three interviews, we would find the probability of success using  $n = 3$  and  $p = 0.40$ . The results are shown in the following table:

<b>Possible Positive Ratings (3 interviews)</b>	0	1	2	3
<b>Theoretical frequencies</b>	21.6	43.2	28.8	6.4

Now we can use the theoretical binomial probabilities of the outcomes to compute the expected frequencies. By comparing these expected frequencies with our observed frequencies using the  $\chi^2$  test, we can examine the extent of the difference between them.

Stating the null and alternative hypotheses

$H_0$ : A binomial distribution with  $p = 0.40$  is a suitable model.

$H_1$ : A binomial distribution with  $p = 0.40$  is not a suitable model.

**Example 5**

The following observations are thought to follow a Poisson distribution.

<b>x</b>	0	1	2	3	4	5	6 or more
<b>f</b>	19	26	27	13	11	2	0

Test whether this model is suitable. Use a 5% significance level.

**Solution 5**

$H_0$ : The Poisson distribution is a suitable model.

$H_1$ : The Poisson distribution is not a suitable model.

$$\bar{x} = \frac{173}{98} = 1.765 \quad \text{Under } H_0: X \sim Po(1.765)$$

Calculating the theoretical probabilities:

$$P(X = x_i) = \frac{e^{-1.765}(1.765)^{x_i}}{x_i!}$$

<b>x</b>	0	1	2	3	4	5 or more
<b>Exp f</b>	16.78	29.61	26.13	15.37	6.78	3.33

Since the last expected frequency is less than 5, combining the last two classes gives

<b>x</b>	0	1	2	3	4 or more
<b>Exp f</b>	16.78	29.61	26.13	15.37	10.11

Reject  $H_0$  if  $\chi^2_{\text{statistic}} > \chi^2_{(\alpha = 0.05)(v = 3)} = 7.81$ .

$$\chi^2_{\text{statistic}} = \frac{(19 - 16.78)^2}{16.78} + \dots + \frac{(13 - 10.11)^2}{10.11} = 1.954$$

Do not reject  $H_0$ : The Poisson distribution is a good fit.

## 254 The Chi-square Test

The goodness of fit test for a normal distribution follows along the same lines as those for a binomial or Poisson. However, in general, given a frequency distribution to test for normality, both the sample mean,  $\bar{x}$  and  $s$ , together with the total frequency, will need to be used. That is, we will generally have *three restrictions* on the choice of the expected values.

The procedure is:

- Calculate  $\bar{x}$  and  $s$  for the given frequency distribution.
- Using  $\bar{x}$  and  $s$  as estimates of the population and  $\sigma$ , together with the given total frequency, set up a theoretical normal distribution.
- Compare observed and expected frequencies in the usual way, using the  $\chi^2$  test statistic with three restrictions.

### Example 6

The following information relates to the height (measured to the nearest cm) of 694 nine-year old girls:

Height	117–120	121–124	125–128	129–132	133–136	137–140	141–144	145–148	149–152
Frequency	8	28	82	140	188	148	69	15	16

Test this sample for normality.

### Solution 6

Calculating from the sample estimates for the population mean and standard deviation gives

$$\bar{x} = 134.356 \quad s = 6.195$$

To calculate the expected frequencies under  $H_0$ : The height of nine-year old girls is normally distributed against  $H_1$ : The height of nine-year old girls is not normally distributed.

The following table sets out the procedure used:

Upper Class Boundary ( $x$ )	$z = \frac{x - 134.356}{6.195}$	$P(Z < z)$	$p$	$E$ ( $p \times 694$ )	$O$
120.5	-2.24	0.013	0.013	9.0	8
124.5	-1.59	0.056	0.043	29.9	28
128.5	-0.95	0.171	0.115	79.8	82
132.5	-0.30	0.382	0.211	146.4	140
136.5	0.35	0.637	0.255	177.0	188
140.5	0.99	0.839	0.202	140.2	148
144.5	1.64	0.950	0.111	77.0	69
148.5	2.28	0.989	0.039	27.1	15
-	-	1.000	0.011	7.6	16

$$\chi^2_{\text{statistic}} = \frac{(8 - 9.0)^2}{9.0} + \frac{(28 - 29.9)^2}{29.9} + \dots + \frac{(16 - 7.6)^2}{7.6} = 17.21$$

Using a 5% significance level the  $\chi^2$  critical value for  $9 - 3 = 6$  degrees of freedom is 12.59.

The result of 17.21 is significant and the assumption of normality,  $H_0$ , is rejected.

### Example 7

A six-face die numbered 1 to 6 is tossed 120 times and the following outcomes were obtained:

Number shown on face	1	2	3	4	5	6
Frequency	19	17	14	18	23	29

Carry out an appropriate test to determine whether the die is biased.

**Solution 7**

The distribution for the number of outcomes may be modelled by a uniform discrete distribution.

$$H_0: \text{The die is fair} - p = \frac{1}{6}$$

$$H_1: \text{The die is not fair} - p \neq \frac{1}{6}$$

For a 5% significance level reject  $H_0$  if

$$\chi^2_{\text{statistic}} > \chi^2_{(\alpha = 0.05)(v = 5)} = 11.07$$

<b>Number shown on face</b>	1	2	3	4	5	6
<b>Observed frequency</b>	19	17	14	18	23	29
<b>Expected frequency</b>	20	20	20	20	20	20

$$\chi^2_{\text{statistic}} = \frac{(19 - 20)^2 + (17 - 20)^2 + \dots + (29 - 20)^2}{20} = 7.00$$

Since  $\chi^2_{\text{statistic}} < \chi^2_{(\alpha = 0.05)(v = 5)}$ , do not reject  $H_0$  – there is not sufficient evidence to conclude that the die is not fair.

## Exercises

1. A cubical die was thrown 300 times, and the results were as follows:

Score	1	2	3	4	5	6
Frequency	48	51	51	58	54	38

Does these results indicate that the die is biased?

2. 3000 observations were made of a random variable and a theoretical “expected” distribution was calculated using the total frequency, to obtain the following table for comparison:

Variable	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
<i>O</i>	620	530	490	410	370	580
<i>E</i>	500	500	500	500	500	500

Comment on the fit, using a  $\chi^2$  test.

3. At a local library, during a given week, the following numbers of books were borrowed

Day	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
Number issued	204	292	242	283	252	275

Is there reason to believe that more books are generally borrowed on one weekday than on another?

4. A bag contains the following coloured discs: 9 red, 7 white, 5 yellow and 4 blue. A disc is drawn at random, its colour noted and returned to the bag. This process is repeated 175 times. The following table shows the observed frequencies:

Colour	Red	White	Yellow	Blue
Frequency	60	51	43	21

- (a) Use the  $\chi^2$  test at the 5% significance level to determine whether there is close agreement between the observed results and the theoretical results you would expect.
- (b) If the process had been repeated 350 times instead of 175 times with observed frequencies 120, 102, 86, and 42, would your conclusion be the same?

9. A hospital administrator has examined past records from 210 randomly selected 8-hour shifts to determine the frequency with which the hospital treats fractures. The number of days in which 0, 1, 2, 3, 4, 5 or more patients with broken bones were treated was 25, 55, 65, 35, 20, and 10 respectively.

At the 5% level of significance, can we reasonably believe that the incidence of broken bone cases follows a Poisson distribution with  $\lambda = 2$ ?

10. In order to plan how much cash on hand should be available for money transactions, a cambio dealer is interested in seeing if the average demand of a customer is normally distributed. The following information was extracted from recent records.

Cash demand (\$)	0–999	1000–1999	2000 and more
Observed frequency	20	65	25

- (a) Compute the expected frequencies if it is thought that the mean cash demand is \$1500 with standard deviation \$600.
- (b) Compute the chi-squared statistic.
- (c) Stating your null and alternative hypotheses clearly, test, at a 10% significance level, whether the average demand for cash is normally distributed.

11. A statistics student wants to see if it is reasonable to assume that some sales data have been sampled from a normal population before performing a hypothesis test on the mean sales. She collected some sales data, computed  $\bar{x} = 78$  and  $s = 9$ , and tabulated the data as follows:

Sales (\$,000)	$\leq 65$	66–70	71–75	76–80	81–85	$\geq 86$
Number of obs. in each group	10	20	40	50	40	40

At the 0.5 level of significance, does the observed distribution follow a normal distribution?

12. In a survey, the weights of 14-year olds,  $X$  kg, were recorded in a group frequency table. The results were:

Weight (kg)	Frequency
$24.5 < x \leq 29.5$	19
$29.5 < x \leq 34.5$	47
$34.5 < x \leq 39.5$	92
$39.5 < x \leq 44.5$	145
$44.5 < x \leq 49.5$	108
$49.5 < x \leq 54.5$	38
$54.5 < x \leq 59.5$	11

# Index

## A

Acceptance region	207
Addition rule	79
Alternative hypothesis	205–206
Approximation, normal to binomial	161–165
Approximation, normal to Poisson	169–170
Approximation, Poisson to binomial	126–127
Association	242
Average	2

## B

Bar chart	17
Binomial distribution	112
Binomial, approximated by normal	161–165
Binomial, approximated by Poisson	126–127
Binomial, expectation and variance	140–142
Bivariate distribution	000
Box-and-whisker diagram	31
Box plot	18

## C

Cdf	106
Central Limit Theorem	179
Chi-squared distribution	243–244
Chi-squared test for contingency table	242–243
Chi-squared test for goodness of fit	250–256
Class boundary	15
Class mid-point	70
Cluster sampling	7, 11
Coded value	68
Combinations	84–90
Common mean, confidence interval	176–182
Conditional probability	80
Confidence interval	176
Confidence interval and hypothesis test	211
Confidence interval for the mean	195
Confidence interval for mean, using $t$	000
Confidence interval for the proportion	198

Confidence limits	202, 206
Contingency table, chi-squared test	242–243
Continuity correction	161
Continuous data	4
Continuous random variable	136–139
Correlation	261–263
Correlation coefficient	262
Correlation, product-moment	267
Covariance	265
Critical region	297
Critical value	252
Cumulative distribution function, (cdf)	105–107
Cumulative frequency diagram	52–53
Cumulative frequency polygon	27–28

## D

Data, raw	3
Data, discrete	3
Data, continuous	4
Decile	27, 142
Degrees of freedom	224
Dependent variable	261
Discrete data	3
Discrete data, class boundary	13
Discrete data, class limit	13
Discrete random variable	97–99
Discrete random variable, expectation	99–100
Discrete random variable, standard deviation	108
Discrete random variable, variance	101–105
Discrete uniform distribution	127–128
Distribution function	146
Distribution of sample mean	179
Distribution, binomial	112
Distribution, chi-squared	243–244
Distribution, geometric	128–130
Distribution, normal	150
Distribution, Poisson	121–124
Distribution, probability	97