

Unit 29

Chi-Square Goodness-of-Fit Test

Objectives:

- To perform the chi-square hypothesis test concerning proportions corresponding to more than two categories of a qualitative variable
- To perform *the Bonferroni method of multiple comparison* when the null hypothesis in a chi-square goodness-of-fit is rejected

The chi-square goodness-of-fit test is a generalization of the z test about a population proportion. If the manager at a particular manufacturing plant believes that 20% of the packages from an assembly line are underweight, the z test about a proportion could be used to see if there is any evidence against this belief. However, if the manager believes that 20% of the packages are underweight, 30% are of the packages are overweight, and 50% of the packages are an acceptable weight, the z test is not appropriate, since more than two categories are involved. When the null hypothesis concerns proportions corresponding to more than two categories of a qualitative variable, the chi-square goodness-of-fit test is appropriate. If we let k represent the number of categories, then k is the number of population proportions. The null hypothesis in a chi-square goodness-of-fit test states hypothesized values for these population proportions.

The test statistic in the chi-square goodness-of-fit test is

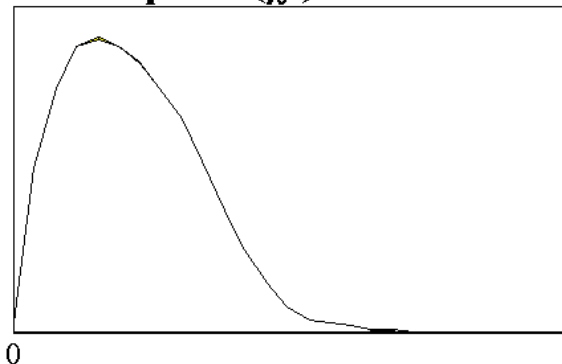
$$\chi^2_{k-1} = \sum \frac{(O - E)^2}{E} ,$$

where O represents observed frequencies and E represents frequencies that are expected if H_0 were actually true. The symbol χ is the Greek letter chi. The superscript 2 on the χ is to emphasize that the test statistic involves squaring, hence χ^2 is read as “chi-squared.” The subscript $k - 1$ refers to the degrees of freedom associated with the test statistic.

In a chi-square goodness-of-fit test, we decide whether or not to reject the null hypothesis by comparing our χ^2 test statistic to an appropriate χ^2 *distribution*. In order to perform the chi-square goodness-of-fit test, we must first study the χ^2 distributions. Table A.5 provides us with information about the χ^2 distributions. The figure at the top of the first page of Table A.5 indicates that each χ^2 distribution, like each f distribution, is positively skewed; also, the values from a χ^2 distribution are nonnegative just as are the values from an f distribution are nonnegative.

Table A.5 contains information about several different χ^2 distributions, each of which is identified by df (degrees of freedom). The information about χ^2 distributions in Table A.5 is organized similar to how the information about t distributions is organized in Table A.3. The rows are labeled with df , and the columns are labeled with various areas under a χ^2 density curve. The notation we shall use to represent the χ^2 -scores in Table A.4 is similar to the notation we use to represent f -scores in Table A.4 and t -scores in Table A.3. For instance, just as we use $t_{5,0.01}$ to represent the t -score above which lies 0.01 of the area under the density curve for the Student's t distribution with $df = 5$, and just as we use $f_{5, 13; 0.01}$ to represent the f -score above which lies 0.01 of the area under the density curve for the Fisher's f distribution with numerator $df = 5$ and denominator $df = 13$, we shall use $\chi^2_{5; 0.01}$ to represent the χ^2 -score above which lies 0.01 of the area under the density curve for the χ^2 distribution with $df = 5$. Table A.5 extends over two

Figure 29-1
Typical Density Curve for a Chi-Square (χ^2) Distribution



pages because unlike the t distributions, the χ^2 distributions are not symmetric. The t distributions have the property that the t -scores in the negative half and the t -scores in the positive half are mirror images of each other; consequently, we only need the information about the positive half. Since the values from a χ^2 distribution are only nonnegative, we must table the values at the left end of the distribution (nearer to zero) and the values at the right end of the distribution separately. For our purposes, we shall only need the values on the first page (the upper end).

Now, let us suppose we collect data for a chi-square goodness-of-fit test, that is, that is, we have observed frequencies for each of k categories of a qualitative variable. If each observed frequency, denoted O , were exactly equal to the corresponding frequency that is expected if H_0 were actually true, denoted by E , then the χ^2 test statistic would be exactly equal to zero. However, even if H_0 were actually true, we still expect there to be some random variation between O and E . The figure at the top of the first page of Table A.5 illustrates the distribution of the χ^2 test statistic if the null hypothesis is true. Figure 29-1 displays this same typical density curve for a χ^2 distribution, and this looks very similar to Figure 27-1 which displays a typical density curve for an f distribution. Notice that, like the typical density curve for an f distribution, the typical χ^2 density curve has a shape which suggests that a high proportion of the χ^2 -scores are roughly speaking in the range from 0 to 2. In a chi-square goodness-of-fit test, a χ^2 test statistic which is unusually large would provide us with evidence that the hypothesized proportions are not all correct. We then define the rejection region in a chi-square goodness-of-fit test by the χ^2 -score above which lies α of the area under the corresponding density curve, where α is of course the significance level. The shaded area in the figure at the top of the first page of Table A.5 graphically illustrates the type of rejection region we use in a chi-square goodness-of-fit test, and we find this to be similar to the type of rejection region in a one-way ANOVA f test.

The same four steps we have used in the past to perform hypothesis tests are used in a chi-square goodness-of-fit test. Let us use a chi-square goodness-of-fit test to see if there is any evidence against the belief that 20% of the packages from an assembly line are underweight, 30% are of the packages are overweight, and 50% of the packages are acceptable weight. We choose a 0.05 significance level to perform this hypothesis test, which is described in Table 29-1a.

Letting λ_U , λ_O , and λ_A respectively represent the population proportions of packages which are underweight, overweight, and an acceptable weight, we complete the first step in our hypothesis test as follows:

$$H_0: \lambda_U = 0.2, \lambda_O = 0.3, \lambda_A = 0.5 \text{ vs. } H_1: \text{Not all of the hypothesized proportions are correct } (\alpha = 0.05).$$

Copy these hypotheses and the significance level in Step 1 of Table 29-1a.

The second step is to collect data and calculate the value of the test statistic. In the simple random sample of 415 packages selected from the assembly line in Table 29-1a, 106 are found to be underweight, 111 are found to be overweight, and 198 are found to have an acceptable weight; these are the observed frequencies, represented by O . If H_0 were true, we would expect that out of 415 randomly selected packages, $(0.2)(415)$ will be underweight, $(0.3)(415)$ will be overweight, and $(0.5)(415)$ will be an acceptable weight. These are the expected frequencies, represented by E . Calculate these expected frequencies, and enter the results in the appropriate places in Step 2 of Table 29-1a; notice that the observed frequencies have already been entered. (You should find that the expected frequencies are respectively 83.0, 124.5, and 207.5.)

Now that we have the values for O and E , it is a simple matter to calculate the test statistic. Before calculating the test statistic, however, a comment about expected frequencies is in order. We do not interpret an expected frequency as a frequency that we must necessarily observe if H_0 is true, because we expect that there will be some random variation. For instance, if you were to flip a perfectly fair coin 10 times, the expected frequency of heads is 5; however, because of random variation, we certainly should not be surprised to observe 6 heads or 4 heads. The correct interpretation of an expected frequency is that it is an average of the frequencies that we would observe if simple random samples of the same size were repeatedly selected. Depending on the sample size, it may not even be possible to actually observe an expected frequency. If you were to flip the fair coin 9 times, the expected frequency of heads is 4.5, which is not possible to actually observe, but 4.5 would be the long term average for expected frequencies.

Recall that the z test about a proportion is appropriate only if the random sample size is sufficiently large; this is also true for the chi-square goodness-of-fit test. In general, the random sample size will be sufficiently large if each expected frequency (E) is greater than or equal to 5. If one or more of the expected

frequencies is less than 5, categories can be combined in a way to insure that all $E \geq 5$. You should notice that each of the expected frequencies calculated in Table 29-1a is greater than 5.

We now return to the calculation of our test statistic. Since our hypothesis test involves $k = 3$ categories, there are $k - 1 = 3 - 1 = 2$ degrees of freedom associated with the test statistic. Substituting the observed and expected frequencies into the test statistic formula, we have

$$\chi^2_2 = \frac{(106 - 83)^2}{83} + \frac{(111 - 124.5)^2}{124.5} + \frac{(198 - 207.5)^2}{207.5} .$$

Do this calculation, and enter the result in the appropriate place in Step 2 of Table 29-1a. (You should find that $\chi^2_2 = 8.272$.) In general, calculating the chi-square goodness-of-fit test statistic is not too difficult, but the appropriate statistical software or programmable calculator will be readily available.

Table 29-1a

Hypothesis Test Concerning Assembly Line Packages

A 0.05 significance level is chosen for a hypothesis test to see if there is any evidence against the belief that 20% of the packages from an assembly line are underweight, 30% are of the packages are overweight, and 50% of the packages are acceptable weight. In a simple random sample of 415 packages selected from the assembly line, 106 are found to be underweight, 111 are found to be overweight, and 198 are found to have an acceptable weight

Step 1 H_0 :

$\alpha =$

H_1 :

Step 2

	Underweight	Overweight	Acceptable
Observed Frequencies (O)	106	111	198
Expected Frequencies (E)	_____	_____	_____

$\chi^2_2 =$

Step 3

Step 4

Table 29-1b

Hypothesis Test Concerning Assembly Line Packages

Bonferroni Method:

	Underweight	Overweight	Acceptable
Sample Proportion	_____	_____	_____
z-score	_____	_____	_____
$z_{\frac{\alpha}{2k}} = z =$	_____		

Statement of results:

The third step is to define the rejection region, decide whether or not to reject the null hypothesis, and obtain the p -value of the test. The shaded area in the figure at the top of the first page of Table A.5 graphically illustrates our rejection region, which is defined by the χ^2 -score above which lies 0.01 of the area under the density curve for the χ^2 distribution with $df = 2$. From Table A.5, find the appropriate χ^2 -score, and then define the rejection region algebraically in Step 3 of Table 29-1a. (You should find the rejection region to be defined by $\chi^2_2 \geq 5.991$.)

Since our test statistic, calculated in the second step, was found to be $\chi^2_2 = 8.272$, which is in the rejection region, our decision is to reject H_0 : $\lambda_U = 0.2$, $\lambda_O = 0.3$, $\lambda_A = 0.5$; in other words, our data provides sufficient evidence that not all of the hypothesized proportions are correct.

The shaded area in the figure at the top of the first page of Table A.5 graphically illustrates the p -value, which is the area above our observed test statistic value $\chi^2_2 = 8.272$ under the density curve for the χ^2 distribution with $df = 2$. This shaded area would represent the probability of obtaining a test statistic value χ^2_2 which represents greater differences between the observed and expected frequencies than the observed test statistic value $\chi^2_2 = 8.272$. By looking at the entries in Table A.5 corresponding to $df = 2$, we find that the observed test statistic $\chi^2_2 = 8.272$ is between $\chi^2_{2; 0.025} = 7.378$ and $\chi^2_{2; 0.01} = 9.210$. This tells us that the p -value is between 0.01 and 0.025, which we can designate by writing $0.01 < p\text{-value} < 0.025$. The fact that $0.01 < p\text{-value} < 0.025$ confirms to us that H_0 is rejected with $\alpha = 0.05$. However, this also tells us that H_0 would not be rejected with $\alpha = 0.01$ but would of course be rejected with $\alpha = 0.10$. Now, complete Step 3 of Table 29-1a.

To complete the fourth step of the hypothesis test, write a summary of the results of the hypothesis test in Step 4 of Table 29-1a. Your summary should be similar to the following:

Since $\chi^2_2 = 8.272$ and $\chi^2_{2; 0.05} = 5.991$, we have sufficient evidence to reject H_0 . We conclude that not all of the hypothesized proportions, $\lambda_U = 0.2$, $\lambda_O = 0.3$, $\lambda_A = 0.5$, are correct ($0.01 < p\text{-value} < 0.025$). Since H_0 is rejected, we need to use multiple comparison to identify which hypothesized proportions are not correct.

When we do not reject the null hypothesis in a chi-square goodness-of-fit test, no further analysis is called for, since we are concluding that the hypothesized proportions are correct. However, rejecting the null hypothesis in a chi-square goodness-of-fit test prompts us to do some further analysis in order to identify which hypothesized proportions are not correct (as we indicated in the summary of results). Which hypothesized proportions are not correct can be identified with multiple comparison. We have previously introduced Scheffe's method of multiple comparison of means. We shall now utilize the *Bonferroni method* to identify which hypothesized proportions are not correct.

Table 29-2

Bonferroni's Multiple Comparison Method to Identify Incorrect Hypothesized Proportions

- (1) Obtain the k sample proportions $\bar{p}_1, \bar{p}_2, \dots, \bar{p}_k$, and the corresponding z -scores
$$\frac{\bar{p}_i - \lambda_{0i}}{\sqrt{\frac{\lambda_{0i}(1 - \lambda_{0i})}{n}}}$$
 for $i = 1, 2, \dots, k$, where $\lambda_{01}, \lambda_{02}, \dots, \lambda_{0k}$ are the respective hypothesized proportions.
- (2) Compare each z -score to the two-sided rejection region defined by
$$z \leq -z_{\frac{\alpha}{2k}} \quad \text{OR} \quad z \geq z_{\frac{\alpha}{2k}}$$
- (3) State the significance level chosen and the direction of statistically significant differences between hypothesized proportions and sample proportions.

Recall that when making multiple comparisons, we want to maintain an overall significance level. With the Bonferroni method, we accomplish this by first dividing the desired overall significance level by $2k$; we then use the resulting significance level on individual comparisons. There are three steps to using the Bonferroni method for identifying which hypothesized proportions are not correct. The first step is to obtain a z -score for each sample proportion, based on the hypothesized proportions. The second step is to obtain a value against which the calculated z -scores are compared, and to identify the hypothesized proportion(s) significantly different from the corresponding hypothesized proportion(s). The third and last step is to state the results. Table 29-2 displays in detail these three steps for the Bonferroni method to identify which hypothesized proportions are not correct.

To illustrate this Bonferroni method, let us identify which hypothesized proportions are not correct from the hypothesis test in Table 29-1. The first step is to obtain a z -score for each sample proportion. Each z -score indicated by the formula in Step 1 of Table 29-2 is of the familiar form where a hypothesized proportion is subtracted from a sample proportion, and the result is divided by a standard error which is the square root of a ratio computed by multiplying the hypothesized proportion and one minus the hypothesized proportion, and dividing this result by the sample size.

For the data in Table 29-1a, you should see that the sample proportion of underweight packages is 106/415, the sample proportion of overweight packages is 111/415, and the sample proportion of packages of acceptable weight is 198/415. Calculate each of these sample proportions, and enter the results in the appropriate places in Table 29-1b. (You should find that the respective sample proportions are 0.2554, 0.2675, and 0.4771.)

Since the sample proportion of underweight packages was found to be 0.2554, and the hypothesized proportion of underweight packages is 0.2, the z -score for the sample proportion of underweight packages is

$$z_U = \frac{0.2554 - 0.2}{\sqrt{\frac{(0.2)(1 - 0.2)}{415}}}$$

Calculate this z -score, and enter the result in the appropriate place in Table 29-1b; then, repeat this for the overweight packages and for the packages of acceptable weight. (You should find that the respective z -scores are +2.821, -1.446, and -0.933.)

The second step of Bonferroni's method is to obtain a value against which the calculated z -scores are compared. We shall use a 0.05 significance level, which is the same significance level chosen in the hypothesis test of Table 29-1a. If we were concerned with whether or not only one sample proportion was different from a hypothesized proportion with $\alpha = 0.05$, we

would in essence be concerned with a two-sided one sample z test. We would compare a single z -score against $z_{0.025}$ (where $\alpha = 0.05$ has been divided by 2, so that we can identify a significant difference in either direction.) However, since we are actually concerned with three proportions, Bonferroni's method has us compare a single z -score against $z_{\alpha / (2k)} = z_{0.05 / (2 \times 3)} = z_{0.05 / 6} = z_{0.0083}$. Step 2 of Table 29-2 tells us that the difference between a sample proportion and the corresponding hypothesized proportion is considered statistically significant when the corresponding z -score is found in the two-sided rejection region defined by

$$z \leq -z_{\alpha / (2k)} \text{ or } z \geq z_{\alpha / (2k)} \text{ (i.e., } |z| \geq z_{\alpha / (2k)} \text{)} .$$

In other words, with the Bonferroni method, we divide the desired overall significance level, which is $\alpha = 0.05$ in the present illustration, by $2k$, which is $(2)(3) = 6$ in the present illustration. We then use the resulting significance level $0.05/6 = 0.0083$ to define the rejection region on each individual comparison. You will not find $z_{0.0083}$ at the bottom of Table A.2 nor at the bottom of Table A.3. To obtain $z_{0.0083}$, you must search through the areas listed in the main body of Table A.2 for the closest area to 0.0083, and obtain the corresponding z -score. Use Table A.2 to verify that $z_{0.0083} = 2.395$, and enter this value in the appropriate place in Table 29-1b.

Now, treating each calculated z -score just like a test statistic in a two-sided z test, look for the significant differences by looking for each calculated z -score which falls into the two-sided rejection region

$$z \leq -2.395 \text{ or } z \geq +2.395 \text{ (i.e., } |z| \geq 2.395 \text{)} .$$

Circle each calculated z -score in Table 29-2b which represents a statistically significant difference. (You should find that the difference between the sample proportion and the hypothesized proportion is statistically significant only for the underweight packages.)

The third and last step of Bonferroni's method is to state the results, including the significance level and the directions of differences. After checking the direction of the significant differences, write your statement of results in the appropriate place in Table 29-1b. Your summary the results of Bonferroni's method should be similar to the following:

With $\alpha = 0.05$, we conclude that the proportion of underweight packages is larger than the hypothesized 0.2.

A graphical display of the data is always helpful in summarizing results. A bar chart or pie chart displaying the sample proportions for each category would be appropriate for the data of Table 29-1a, since the data consists of the one qualitative variable "type of package" with the three categories. Figure 29-1 is a pie

Figure 29-1

Proportions for Selected Packages

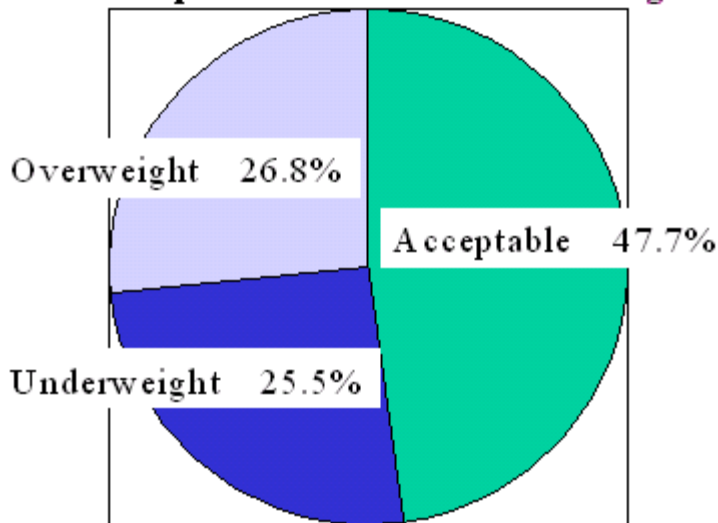


chart displaying the sample proportions for underweight packages, overweight packages, and packages of acceptable weight.

It should be no surprise that applying a chi-square goodness-of-fit test when there are only $k = 2$ categories will produce exactly the same results as the one sample z test about the proportion for one of the two categories. In fact, each of the χ^2 -scores with 1 degree of freedom in Table A.5 is the square of a corresponding z -score. For instance, you can check that $[z_{0.025}]^2 = 1.960^2 = 3.841$, and that $\chi^2_{1;0.05} = 3.841$.

Self-Test Problem 29-1. A particular car manufacturer has believed for several years that among buyers of a particular type of sports car, 40% prefer red, 30% prefer green, 20% prefer yellow, and 10% prefer white. In a random sample of 252 buyers of the sports car, 114 prefer red, 57 prefer green, 42 prefer yellow, and 39 prefer white. A 0.05 significance level is chosen for a hypothesis test to see if there is any evidence against this belief.

- (a) Explain how the data for this hypothesis test is appropriate for a chi-square goodness-of-fit test.
- (b) Complete the four steps of the hypothesis test by completing the table titled *Hypothesis Test for Self-Test Problem 29-1*. You should find that $\chi^2_3 = 15.262$.
- (c) If multiple comparison is necessary, apply Bonferroni's method and state the results; if multiple comparison is not necessary, explain why not.
- (d) Verify that the sample size is sufficiently large for the chi-square goodness-of-fit test to be appropriate.
- (e) Considering the results of the hypothesis test, decide which of the Type I or Type II errors is possible, and describe this error.
- (f) Decide whether H_0 would have been rejected or would not have been rejected with each of the following significance levels: (i) $\alpha = 0.01$, (ii) $\alpha = 0.10$.
- (g) Construct an appropriate graphical display for the data used in this hypothesis test.

Hypothesis Test for Self Test Problem 29-1

Step 1 H_0 :

H_1 :

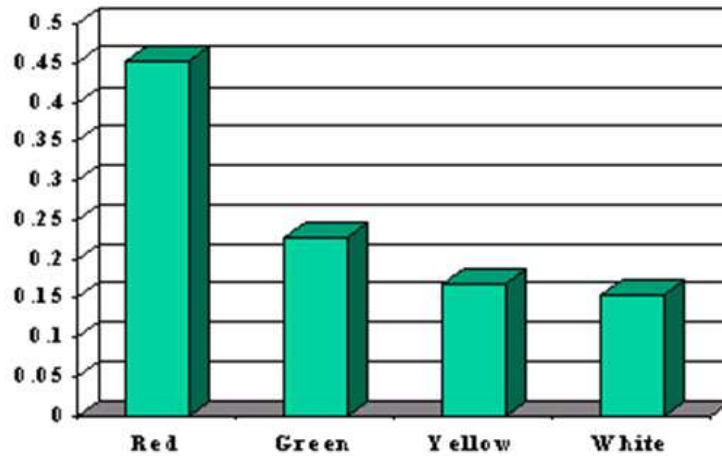
$\alpha =$

Step 2

Step 3

Step 4

Figure 29-3
Bar Chart of the Data for Self-Test Problem 29-1



Answers to Self-Test Problems

29-1 (a) The data consists of a random sample of observations of the qualitative variable “color,” and the purpose of a chi-square goodness-of-fit test is to compare the sample category proportions of the qualitative variable to hypothesized proportions.

(b) Step 1: $H_0: \lambda_R = 0.4, \lambda_G = 0.3, \lambda_Y = 0.2, \lambda_W = 0.1$ ($\alpha = 0.05$)

H_1 : At least one hypothesized proportion is not correct

Step 2: For the respective colors red, green, yellow, and white, the expected frequencies are 100.8, 75.6, 50.4, and 25.2. $\chi^2_3 = 15.262$

Step 3: The rejection is $\chi^2_3 \geq 7.815$. H_0 is rejected; $0.001 < p\text{-value} < 0.005$.

Step 4: Since $\chi^2_3 = 15.262$ and $\chi^2_{3; 0.05} = 7.815$, we have sufficient evidence to reject H_0 . We conclude that not all of the hypothesized proportions, $\lambda_R = 0.4, \lambda_G = 0.3, \lambda_Y = 0.2, \lambda_W = 0.1$ are correct ($0.001 < p\text{-value} < 0.005$). Since H_0 is rejected, we need to use multiple comparison to identify which hypothesized proportions are not correct.

(c)	<u>Red</u>	<u>Green</u>	<u>Yellow</u>	<u>White</u>
Sample Proportion	114/252 = 0.4524	57/252 = 0.2262	42/252 = 0.1667	39/252 = 0.1548.
z-score	+1.698	-2.557	-1.322	+2.900
	$z_{0.0062} = 2.50$			

With $\alpha = 0.05$, we conclude that the proportion of buyers preferring green is smaller than the hypothesized 0.3, and that the proportion of buyers preferring white is larger than the hypothesized 0.1.

(d) Since the expected frequencies are all greater than 5, the sample size is sufficiently large for chi-square goodness-of-fit test to be appropriate.

(e) Since H_0 is rejected, the Type I error is possible, which is concluding that at least one hypothesized proportion is not correct when in reality the hypothesized proportions are all correct.

(f) H_0 would have been rejected with both $\alpha = 0.01$ and $\alpha = 0.10$.

(g) Since color is a qualitative variable, a bar chart or pie chart is an appropriate graphical display. Figure 29-3 displays a bar chart.

Summary

With a sufficiently large sample size, the chi-square goodness-of-fit hypothesis test concerning proportions corresponding to $k > 2$ categories can be performed by using the chi-square test statistic

$$\chi^2_{k-1} = \sum \frac{(O - E)^2}{E} ,$$

where O observed frequencies and E represents frequencies that are expected if H_0 were actually true. The null hypothesis, which gives hypothesized proportions, is rejected when the chi-square test statistic is larger than a value determined from the *chi-square distribution with $k - 1$ degrees of freedom*. An easy way to check that the sample size is sufficiently large is to verify that all $E \geq 5$.

When we do not reject the null hypothesis in a chi-square goodness-of-fit test, no further analysis is called for, since we are concluding that the hypothesized proportions are correct. When we reject the null hypothesis in a chi-square goodness-of-fit test, multiple comparison is desirable to identify which hypothesized proportions are not correct.

The *Bonferroni method* is one of many multiple comparison methods available. When applied to the sample proportions from a chi-square goodness-of-fit test, this method consists of the three steps displayed in Table 29.2.