Chi-Square—Chapter 11, Sections 11.1-11.2

Your text discusses three types of chi-square tests:
1. Goodness-of-Fit
2. Independence
3. Homogeneity of Proportions

We use the Greek symbol χ^2 (pronounced "$k\bar{i} - square$" to rhyme with "sky-square") to represent values of the chi-square distribution.

Characteristics of the Chi-Square Distribution

1. It is not symmetric.
2. The values of χ^2 are non-negative (i.e. $\chi^2 > 0$).
3. The chi-square distribution is asymptotic to the horizontal axis on the right-hand-side.
4. The shape of the chi-square distribution depends upon the degrees of freedom, just like Student's t-distribution and Fisher's F-distribution.
5. As the number of degrees of freedom increases, the chi-square distribution becomes more symmetric, as illustrated in Figure 1 below.
6. Total area under the curve is equal to 1.0.

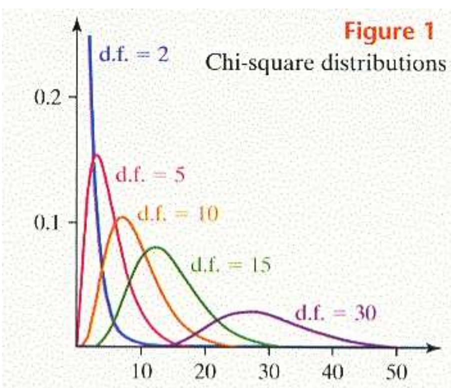


Figure 1
Chi-square distributions

Finding Critical Values of the Chi-Square Distribution:

Critical values of the chi-square distribution are found in Table IV in Appendix A.

Find the critical value of chi-square for a one-tail (right-tail) test with $\alpha = 0.05$ and df=15.

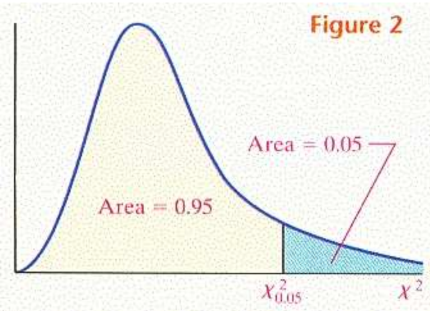


Figure 2

Figure 3

Degrees of Freedom	Area to the Right of the Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
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	11.071	12.833	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
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.365	7.215	8.231	9.288	10.265	25.200	28.601	31.595	24.265	36.456

11.1 Chi-Square Goodness-of-Fit Test

Definition: A **goodness-of-fit test** is an inferential procedure used to determine whether a frequency distribution follows a claimed distribution. It is a test of the agreement or conformity between the observed frequencies (O_i) and the expected frequencies (E_i) for several classes or categories

Expected Frequencies (or Counts)

Suppose there are n independent trials of an experiment with $k \geq 3$ mutually exclusive possible outcomes. Let p_1 represent the probability of observing the first outcome and E_1 represent the expected frequency of the first outcome; p_2 represent the probability of observing the second outcome and E_2 represent the expected frequency of the second outcome; and so on. The expected frequencies for each possible outcome are given by

$$E_i = np_i \qquad \text{for } i=1, 2, \ldots, k.$$

Theorem: **Test Statistic for Goodness-of-Fit Tests**

Let O_i represent the observed counts of category i; E_i represent the expected counts of category i; k represent the number of categories; and n represent the number of independent trials of an experiment. Then the formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad i = 1, 2, \ldots, k$$

approximately follows the chi-square distribution with k-1 degrees of freedom, provided that (1) all expected frequencies are greater than or equal to 1 (all $E_i \geq 1$) and (2) no more than 20% of the expected frequencies are less than 5 (where $E_i=np_i$ for i=1, 2, …, k).

Note: If $O_i=E_i$ for all i, then $\chi^2 = 0$. >>>Perfect agreement between observed and expected frequencies.

If O_i differs from E_i, then $\chi^2 > 0$ and increases in size as the difference increases.

The Chi-Square Goodness-of-Fit Test

Assumptions:

- The data are randomly selected.
- All expected frequencies are greater than or equal to 1 (i.e., $E_i \geq 1$.)
- No more than 20% of the expected frequencies are less than 5.

Step 1: A claim is made regarding a distribution.

H_0: The random variable follows the claimed distribution.

H_1: The random variable does not follow the claimed distribution.

Step 2: Select a significance level, α, and find the **critical value of chi-square, χ_α^2** with df=k-1
Note that chi-square goodness-of-fit tests are right-tailed tests, so the critical value is always on the right tail.

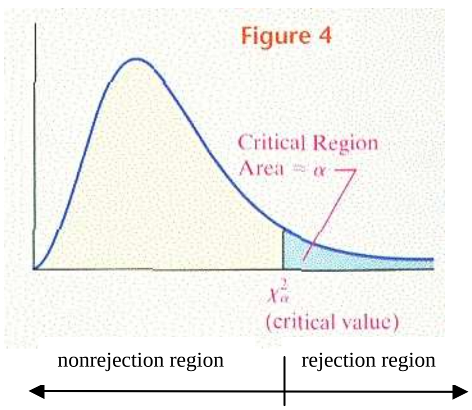


Figure 4

Step 3: Calculate the χ^2 **test-statistic**:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad i = 1, 2, \ldots, k$$

where $E_i = np_i$ for each of the k categories, with n=number of trials and p_i=probability of the ith category. The E_i are derived under the assumption that the null hypothesis is true.

Step 4: Draw a conclusion.

- Compare the test statistic with the critical value. If $\chi^2 > \chi_\alpha^2$, reject H_0.
- Interpret the conclusion in the context of the problem.

Example of Chi-Square Goodness-of-Fit Test:

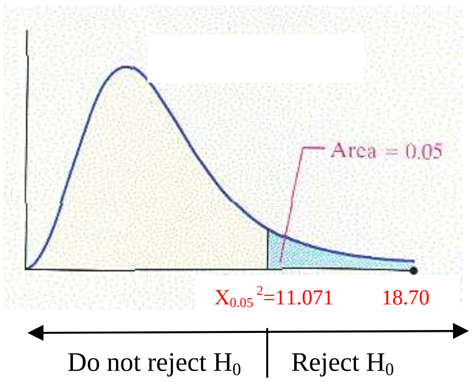
Problem: A die is tossed 120 times. Test the hypothesis that the die is "fair."

Step 1: Null and alternative hypotheses

H_0: Die is "fair" (i.e., $p_1=p_2=\ldots=p_6=1/6$)

H_1: Die is not fair (i.e., at least one $p_i \neq 1/6$)

Step 2: Select α=0.05 and find the critical value of chi-square, $\chi_{0.05}^{\ 2} = 11.071$ with df=(6-1)=5.



Step 3: Toss the die 120 times and record the number of 1's, 2's, …, and 6's. Calculate the expected frequencies using $E_i=np_i$.

No. on Face of Die	O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
1	13	20	$\frac{(13-20)^2}{20} = 2.45$
2	28	20	$\frac{(28-20)^2}{20} = 3.20$
3	16	20	$\frac{(16-20)^2}{20} = 0.80$
4	10	20	$\frac{(10-20)^2}{20} = 5.00$
5	32	20	$\frac{(32-20)^2}{20} = 7.20$
6	21	20	$\frac{(21-20)^2}{20} = 0.05$
	120	120	$\sum\frac{(O_i-E_i)^2}{E_i} = 18.70$

$E_1=120(1/6) = 20$

The H_0 is rejected largely due to the small number of observed 4's (O_4=10) and the large number of observed 5's (O_5=32).

Step 4: Conclusion—Because the χ^2-statistic=18.70 > $\chi_{0.05}^{\ 2}$ = 11.071, reject H_0 at the 0.05 significance level. The sample data imply that the die is not fair.

***Re-work this problem using the P-value approach explained on p. 482 of your text.

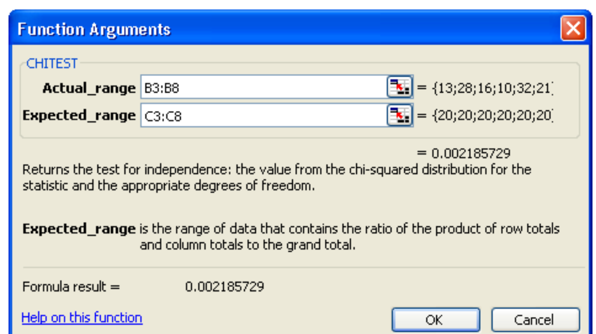
Excel—Calculation of the Test Statistic and P-Value for a Chi-Square Goodness-of-fit Test

Excel can be used for calculations for a chi-square goodness-of-fit test. The Die Toss problem from p. 5 is used as an example. It is important to point out that Excel does not provide a statistical function to calculate the E_i's for a chi-square problem; instead, the E_i's must be calculated with standard Excel functions for add, subtract, multiply, and divide (following the hand-calculation procedures demonstrated on p. 5).
In regard to application of a chi-square statistical test, Excel can provide three pieces of information to facilitate the process:

Measure Calculated in Excel	**Excel Statistical Function**
1. P-value indicating the area under the chi-square distribution to the right of the test statistic.	**CHITEST**
2. Chi-square test statistic, χ^2.	**CHIINV**
3. Chi-square critical value, χ_α^2.	**CHIINV**

The use of Excel in obtaining these measures is demonstrated below for the Die Toss problem from p. 5. The expected frequencies (E_i) are calculated as demonstrated on p. 5 using standard Excel functions for add, subtract, multiply, and divide.

1	A	B	C
2	No. on Face of Die	Observed O_i	Expected E_i
3	1	13	20
4	2	28	20
5	3	16	20
6	4	10	20
7	5	32	20
8	6	21	20
9		120	120
10			
11	CHITEST	0.002186	
12	CHIINV	18.7	
13	CHIINV	11.070	



The **CHITEST** function can be used to calculate the χ^2 p-value.

Excel: Finding P-values for the χ^2 distribution.
Step 1: Calculate the expected frequencies (E_i).
Step 2: Select **Insert/Function (f_x)** from the Windows menu. In the **Function Category**, select "Statistical." In the **Function Name**, select "CHITEST" and press **OK** and the **Function Argument** window will appear.

a) Press the button on the right of the **Actual_range** box and highlight cells **B3:B8**. This highlights the observed frequencies.
b) Press the button on the right of the **Expected_range** box and highlight cells **C3:C8**. This highlights the expected frequencies.

Step 3: Press **OK** and the p-value will appear at the bottom of the **Function Arguments** box.
Formula result = 0.002185729 (this is the P-value)

The CHIINV function can be used to calculate the chi-square test statistic, χ^2, as explained below.

Excel: Finding the chi-square statistic (or calculated chi-square value)

Step 1: Calculate the p-value using the "CHITEST" function (as explained above)

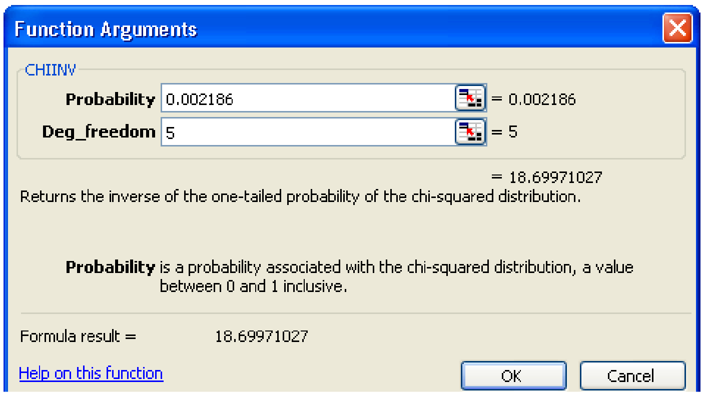
Step 2: Select **Insert/Function** (**f_x**) from the Windows menu. In the **Function Category**, select "Statistical." In the **Function Name**, select "CHIINV" and press **OK** and the **Function Argument** window will appear.

a) Enter the P-value in the **Probability** box (the P-value was calculated using the "CHITEST" function in Step 1).

b) Enter (n-1) in the **Deg_freedom** box (note that df identifies the particular chi-square distribution that is being used).

Step 3: Press **OK** and the chi-square statistic will appear at the bottom of the **Function Arguments** box.

Formula result = 18.69971027



The CHIINV function can be used to find the chi-square critical value, χ_α^2, as explained below

Excel: Finding critical values for the chi-square distribution.

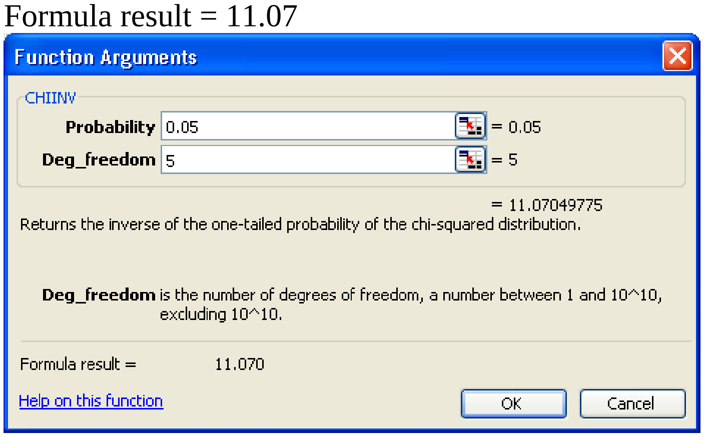
Select **Insert/Function** (**f_x**) from the Windows menu. In the **Function Category**, select "Statistical." In the **Function Name**, select "CHIINV" and press **OK** and the **Function Argument** window will appear.

a) Enter the significance level (α) in the **Probability** box. .

b) Enter the degrees of freedom in the **Deg_freedom** box (df=n-1).

Press **OK** and the chi-square statistic will appear at the bottom of the **Function Arguments** box.

Formula result = 11.07



11.2 Chi-Square Test for Independence

Consider the data in the table below that represent the eye color and hair shade of a random sample of 50 individuals.

	Hair Shade	
Eye Color	**Light Hair**	**Dark Hair**
Blue Eyes	23	7
Brown Eyes	4	16

This table is referred to as a **contingency table** or a **two-way table** because it relates two categories of data. The **row variable** is eye color and the **column variable** is hair shade. Each box in the table is referred to as a **cell**. For example, the cell referring to light hair and blue eyes is in the first row, first column. Each cell contains the frequency of the category.

Definition: The **chi-square independence test** is used to find out whether there is an association between a row variable and column variable in a contingency table constructed from sample data. The null hypothesis is that the variables are not associated: in other words, they are independent. The alternative hypothesis is that the variables are associated, or dependent.

Expected Frequencies in a Chi-Square Independence Test

To find the expected frequencies in a cell when performing a chi-square independence test, multiply the row total of the row containing the cell by the column total of the column containing the cell and divide this result by the table total. That is,

$$Expected\ frequency = \frac{(row\,total)(column\,total)}{table\,total}$$

***Theorem:* Test Statistic for the Test of Independence**

Let O_i represent the observed counts in the ith cell and E_i represent the expected counts in the ith cell. Then

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

approximately follows the chi-square distribution with (r-1)(c-1) degrees of freedom, where r is the number of rows and c is the number of columns in the contingency table, provided that (1) all expected frequencies are greater than or equal to 1 (all $E_i \geq 1$) and (2) no more than 20% of the expected frequencies are less than 5.

The Chi-Square Test for Independence

Assumptions

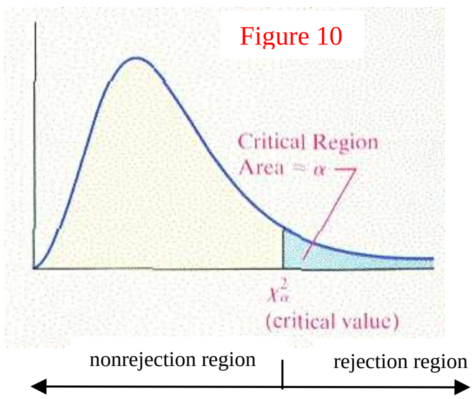
- The data are randomly selected.
- All expected frequencies are greater than or equal to 1 (i.e., $E_i \geq 1$.)
- No more than 20% of the expected frequencies are less than 5.

Step 1: A claim is made regarding the independence (or dependence) of two variables.

H_0: The row variable and column variable are independent.

H_1: The row variable and column variable are dependent.

Step 2: Select a significance level, α, and find the **critical value of chi-square.** All chi-square independence tests are right-tailed tests, so the critical value is χ_α^2 with (r-1)(c-1) degrees of freedom. The shaded region represents the critical region in the figure below.



Step 3: Compute the expected frequencies using the formula: $Exp.\ freq = \frac{(row\,tot)(col\,tot)}{table\,total}$

Compute the **test statistic**:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Step 4: Draw a conclusion.

- Compare the test statistic to the critical chi-square. If $\chi^2 > \chi_\alpha^2$, reject H_0.
- Interpret the conclusion in the context of the problem.

Example of Chi-Square Independence Test:

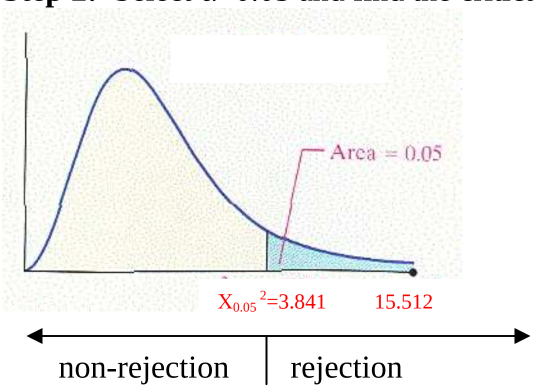
Problem: Is eye color and shade of hair related in individuals? Can we conclude from the data shown below that there is a significant connection between eye color and hair shade?

Step 1: Null and alternative hypotheses

H_0: Eye color and hair shade are independent.

H_1: Eye color and hair shade are dependent.

Step 2: Select $\alpha=0.05$ and find the critical value of $\chi_\alpha^2 = 3.841$ with df=(2-1)(2-1)=1.



Step 3:

A random sample of 50 individuals was selected and classified according to eye color and shade of hair (shown at the top of the table below). Expected frequencies were calculated using the formula: $Exp.\ freq = \frac{(row\,tot)(col\,tot)}{table\,total}$ (shown at the bottom of the table).

Observed Frequencies:

	Light Hair	Dark Hair	***Row Total***
Blue Eyes	23	7	30
Brown Eyes	4	16	20
Column Total	27	23	50

Expected Frequencies:

	Light Hair	Dark Hair	***Row Total***
Blue Eyes	$\frac{(30)(27)}{50}=16.2$	$\frac{(30)(23)}{50}=13.8$	30
Brown Eyes	$\frac{(20)(27)}{50}=10.8$	$\frac{(20)(23)}{50}=9.2$	20
Column Total	27	23	50

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(23-16.2)^2}{16.2} + \frac{(7-13.8)^2}{13.8} + \frac{(4-10.8)^2}{10.8} + \frac{(16-9.2)^2}{9.2} = 15.512$$

Step 4: Conclusion—Because the calculated χ^2=15.521 > the critical χ_α^2 = 3.841, reject H_0 at the 0.05 significance level. A significant relation exists between eye color and hair shade (i.e., eye color and hair shade are dependent).

***Re-work this problem using the P-value approach explained on p. 494 of your text.

Constructing a Conditional Distribution and Bar Graph (p. 493)

Problem: Find the conditional distribution of eye color by hair shade. Then draw a bar graph that represents the conditional distribution of eye color by hair shade.

Table of Conditional Distributions

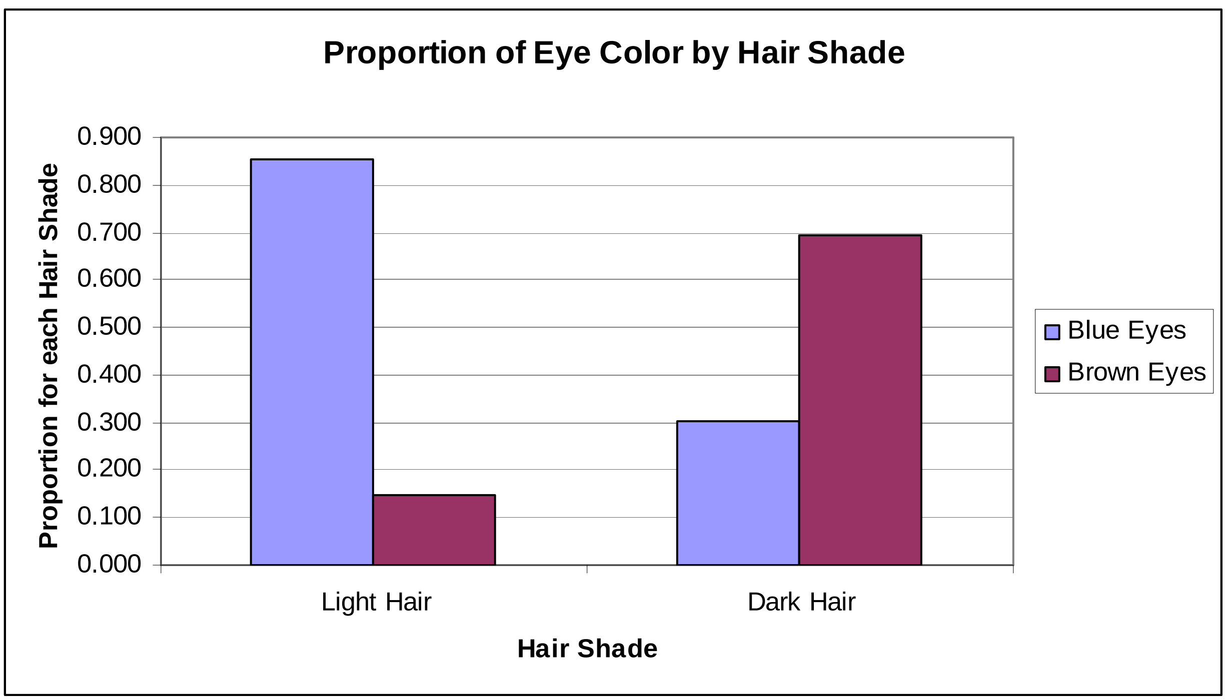
Observed Frequencies:

	Light Hair	**Dark Hair**
Blue Eyes	23	7
Brown Eyes	4	16
Column Total	27	23

Proportions:

	Light Hair	**Dark Hair**
Blue Eyes	$\frac{23}{27} = 0.852$	$\frac{7}{23} = 0.304$
Brown Eyes	$\frac{4}{27} = 0.148$	$\frac{16}{23} = 0.696$
	1.000	1.000

The graph below (drawn with the Excel Chart Wizard) shows the conditional distribution of eye color by hair shade; the blue bars represent the proportion of people with blue eyes for each hair shade, and the maroon bars represent the proportion of people with brown eyes for each hair shade. The association between eye color and hair shade is apparent from these conditional distributions. The proportion of people with blue eyes is greater for those with light hair compared to dark hair (the relation is significant at the 0.05 level).



Chi-Square Test for Homogeneity of Proportions

The chi-square test for independence (that was just discussed above) is a test regarding a sample from a *single* population. Each individual in the population is classified in two ways (such as eye color and hair shade). We now discuss a second type of chi-square test, which can be used to compare the population proportions from *different* populations. The test we are about to introduce is an extension of the procedures introduced in Section 10.3 where we compared two population proportions.

> *Definition:* In a **chi-square test for homogeneity of proportions**, we test the claim that different populations have the same proportion of individuals with some characteristic.

Example: How patriotic are you? Would you say—extremely/very patriotic, somewhat patriotic, or not especially patriotic? This question was asked of a random sample of 1009 adults in June 1999 and in June 2005 (The Gallup Organization, http://www.gallup.com, June 30, 2005). We might initially claim that the proportion of individuals in each category (extremely/very patriotic, somewhat patriotic, not especially patriotic) is the same for the 1999 population as for the 2005 population. In other words, the level of patriotism in the American population has not changed from 1999 to 2005. The appropriate null hypothesis is:

H_0: $p_1 = p_2$,

versus the alternative hypothesis:

H_1: At least one of the population proportions is different from the other.
(e.g., p_1=proportion of people who are extremely/very patriotic in 1999 is different from p_2=proportion of people who are extremely/very patriotic in 2005)

The procedures for performing a test for homogeneity of proportions are identical to those for a test of independence.

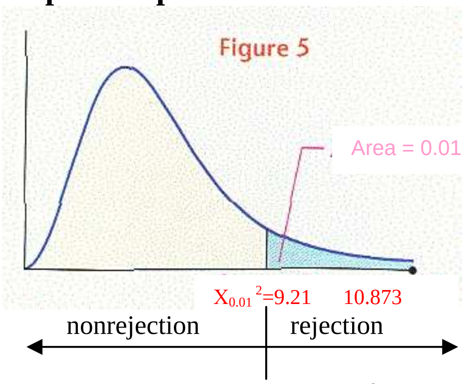
Problem: Is there evidence to indicate that the feeling of adults toward patriotism is the same in 2005 as in 1999? Perform a test for homogeneity of proportions using $\alpha=0.01$.

Step 1: Null and alternative hypotheses:

H_0: $p_1 = p_2$ where 1=extremely/very patriotic in 1999 and 2=extremely/very patriotic in 2005; and so on for the other categories.

H_1: At least one of the population proportions is different from the other.

Step 2: **Step 2:** Select $\alpha=0.01$ and find the critical value of $\chi_\alpha^2 = 9.210$ with df=(3-1)(2-1)=2.



Step 3: A random sample of 1009 adults was asked the question: How patriotic are you? Would you say—extremely/very patriotic, somewhat patriotic, or not especially patriotic? This question was asked in June 1999 and again in June 2005 and the observed frequencies are shown at the top of the table below. The expected frequencies were calculated using the formula:

$Exp.\ freq = \dfrac{(row\,tot)(col\,tot)}{table\,total}$ (expected frequencies are shown at the bottom of the table below).

Observed Frequencies:

	June 1999	June 2005	*Row Total*
Extremely/Very patriotic	656	726	1382
Somewhat patriotic	283	222	505
Not especially patriotic	50	50	100
Column Total	989	998	1987

Note: The column totals differ from 1009 due to "No Opinion" responses.

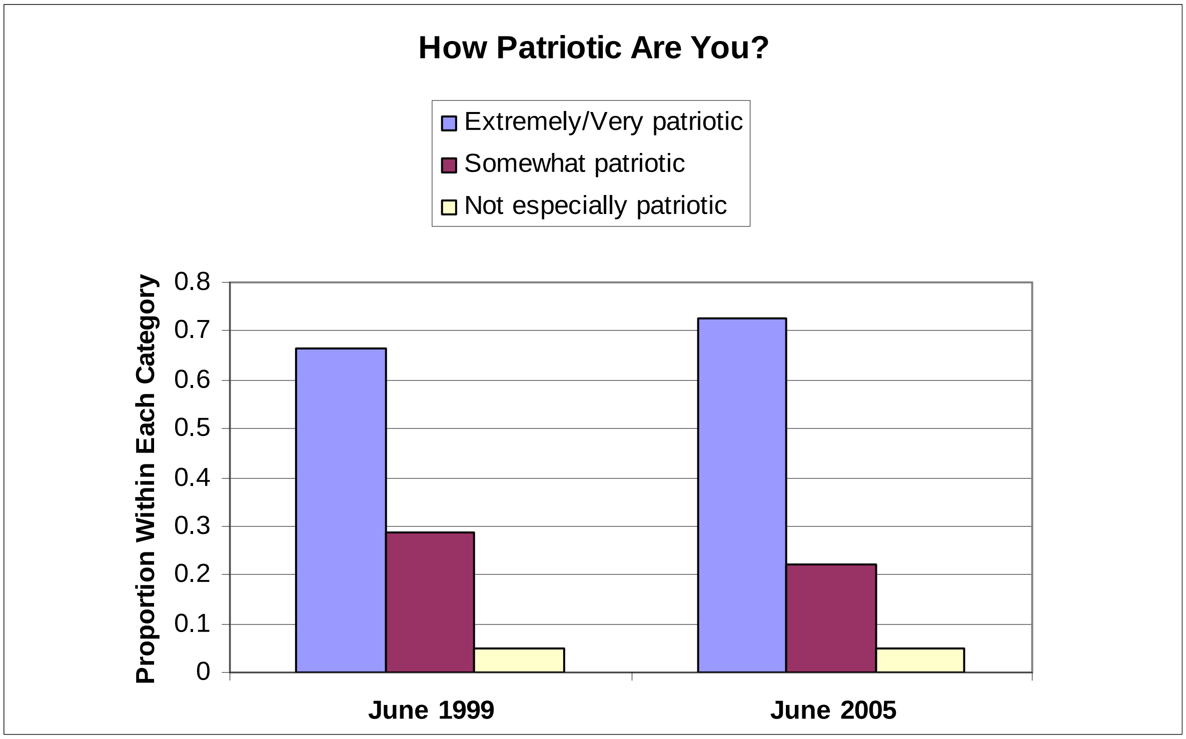
Expected Frequencies:

	June 1999	June 2005	*Row Total*
Extremely/Very patriotic	$\frac{(1382)(989)}{1987} = 687.87$	$\frac{(1382)(998)}{1987} = 694.13$	1382
Somewhat patriotic	$\frac{(505)(989)}{1987} = 251.36$	$\frac{(505)(998)}{1987} = 253.64$	505
Not especially patriotic	$\frac{(100)(989)}{1987} = 49.77$	$\frac{(100)(998)}{1987} = 50.23$	100
Column Total	989	998	1987

$$\chi^2 = \Sigma\frac{(O_i - E_i)^2}{E_i} = \frac{(656-687.87))^2}{687.87} + \frac{(726-694.13)^2}{694.13} + \frac{(283-251.36)^2}{251.36} + \frac{(222-253.64)^2}{253.64} + \frac{(50-49.77)^2}{49.77} + \frac{(50-50.23)^2}{50.23} = 10.873$$

Step 4: Conclusion—Because the calculated χ^2=10.873 > the critical $\chi_\alpha^2 = 9.210$, reject H_0 at the 0.01 significance level. There is a significant difference in the level of patriotism between 1999 and 2005.

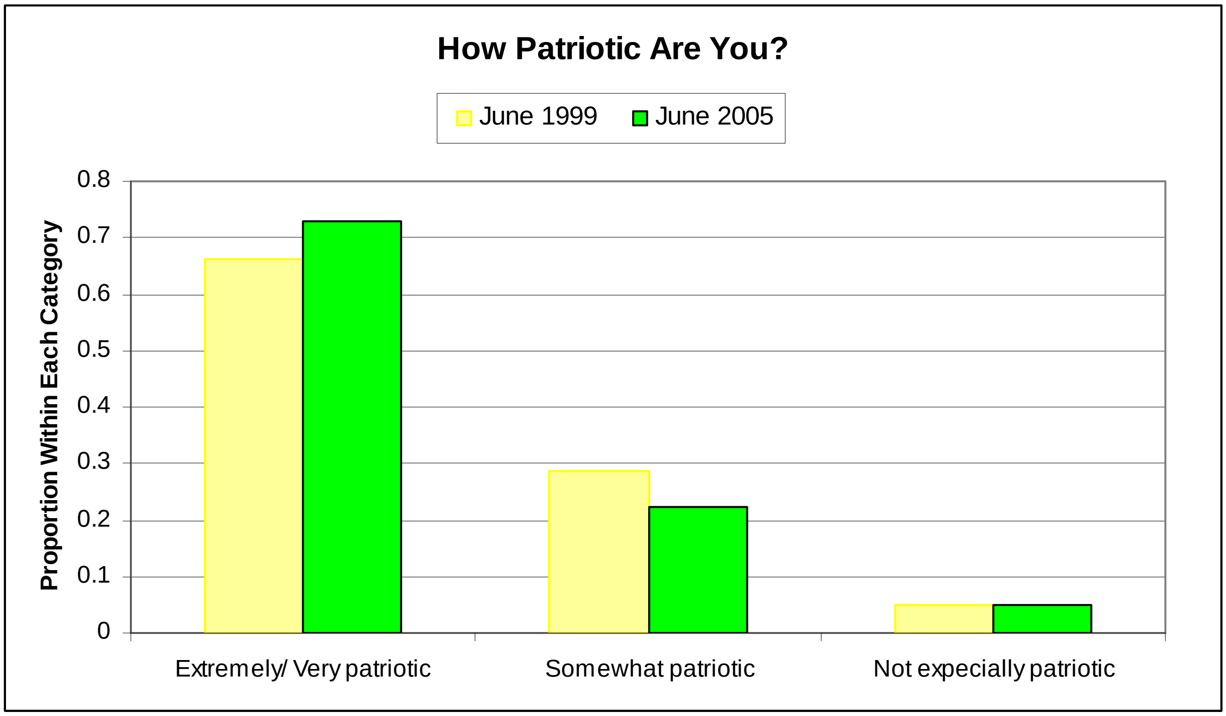
The graph shows a bar graph for the conditional distributions of patriotism level for 1999 and 2005. The blue bars represent the proportion of people who are *extremely/very patriotic*; the maroon bars represent the proportion who are *somewhat patriotic*; and the yellow bars represent the proportion who are *not especially patriotic*. From the graph it is apparent that the proportion of people who are extremely/very patriotic has increased slightly from 1999 to 2005 (the change is significant at the 0.01 level).



Source: The Gallup Organization (http://www.gallup.com), June 30, 2005.

The graph shows a bar graph for the conditional distributions of patriotism level for 1999 and 2005. The yellow bars represent *June 1999* and the green bars represent *June 2005*. From the graph it is apparent that the proportion of people who are extremely/very patriotic has increased slightly from 1999 to 2005 whereas the proportion of people who are somewhat patriotic has decreased slightly (the changes are significant at the 0.01 level).

Observed Frequencies				
	Extremely/Very patriotic	**Somewhat patriotic**	**Not especially patriotic**	**Row Total**
June 1999	656	283	50	989
June 2005	726	222	50	998
Column Total	1382	505	100	1987
Note: The column totals differ from 1009 due to "No Opinion" responses.				
Expected Frequencies				
	Extremely/Very patriotic	**Somewhat patriotic**	**Not especially patriotic**	**Row Total**
June 1999	$\frac{(989)(1382)}{1987} = 687.87$	$\frac{(989)(505)}{1987} = 251.36$	$\frac{(989)(100)}{1987} = 49.77$	989
June 2005	$\frac{(998)(1382)}{1987} = 694.13$	$\frac{(998)(505)}{1987} = 253.64$	$\frac{(998)(100)}{1987} = 50.23$	998
Column Total	1382	505	100	1987



Source: The Gallup Organization (http://www.gallup.com), June 30, 2005.