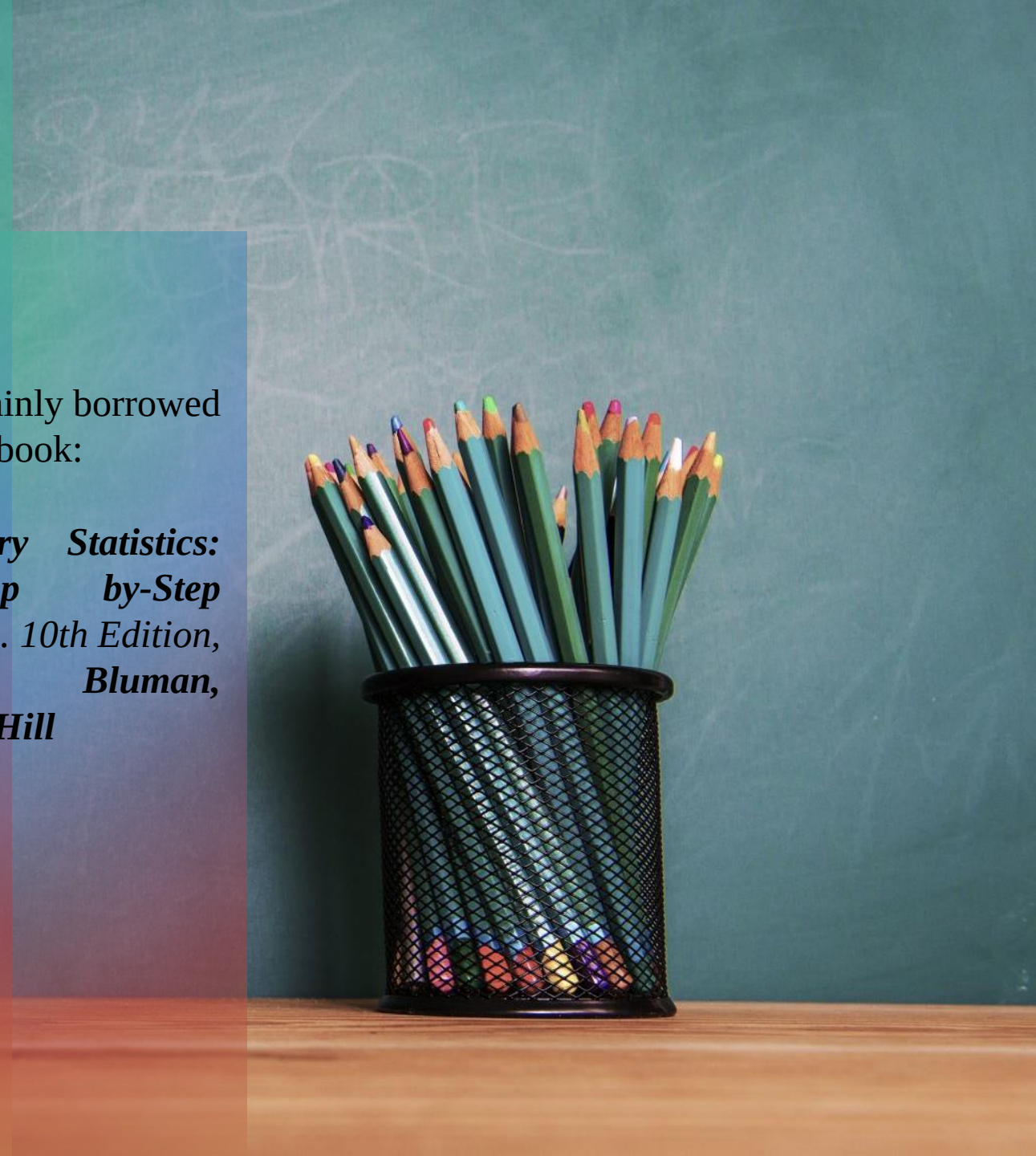
Intermediate Analytics

Fatemeh Ahmadi

ALY 6015

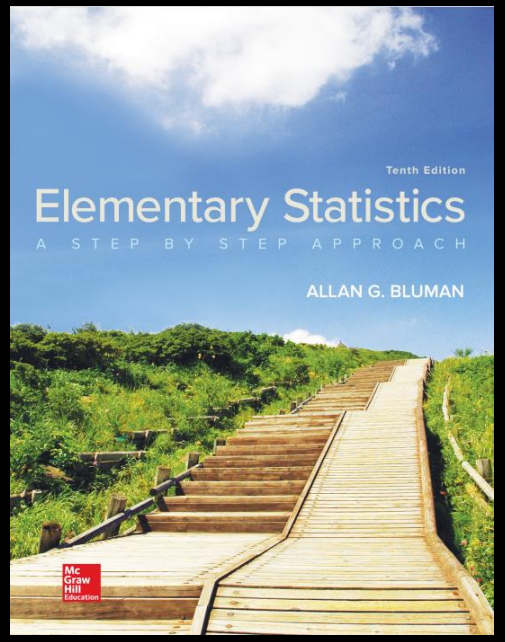
Slides are mainly borrowed from the textbook:

- ***Elementary Statistics: A Step by-Step Approach***. *10th Edition,* ***Allen Bluman, McGraw Hill***



Chi-Square

You will learn in this course:



After completing this chapter, you should be able to:

1. Test a distribution for the goodness of fit, using Chi-Square.
2. Test two variables for independence, using Chi-Square.
3. Test proportions for homogeneity, using Chi-Square.

Introduction

- The Chi-Square can be used for tests concerning **frequency distributions**, such as “If a sample of buyers is given a choice of automobile colors, will each color be selected with the same frequency?”
- The Chi-Square distribution can be used to test the **independence of two variables**, for example, “Are senators’ opinions on gun control independent of party affiliations?” That is, do the republicans feel one way and the democrats feel differently, or do they have the same opinion?
- Finally, the Chi-Square distribution can be used to test the **homogeneity of proportions**. For example, is the proportion of high school seniors who attend college immediately after graduating the same for the northern, southern, eastern, and western parts of the United States?

Test for Goodness of Fit

- In addition to being used to test a single variance, the Chi-Square statistic can be used to see whether **a frequency distribution fits a specific pattern**. For example, to meet customer demands, a manufacturer of running shoes may wish to see whether buyers show a preference for a specific style.

- A traffic engineer may wish to see whether accidents occur more often on some days than on others so that he/she can increase police patrols accordingly.

- An emergency service may want to see whether it receives more calls at certain times of the day than at others so that it can provide adequate staffing.

Test for Goodness of Fit

Recall the characteristics of the Chi-Square distribution:

1. The Chi-Square distribution is a family of curves based on the degrees of freedom.
2. The Chi-Square distributions are positively skewed.
3. All Chi-Square values are greater than or equal to zero.
4. The total area under each Chi-Square distribution is equal to 1.

When you are testing to see whether a frequency distribution fits a specific pattern, you can use the **Chi-Square goodness-of-fit test**.

The **chi-square goodness-of-fit test** is used to test the claim that an observed frequency distribution fits some given expected frequency distribution.

Example 1

For example, suppose you wanted to see if there was a difference in the number of arrests in a certain city for four types of crimes. A random sample of 160 arrests showed the following distribution.

Larceny thefts	Property crimes	Drug use	Driving under the influence
38	50	28	44

- ✓ Since the frequencies for each flavor were obtained from a sample, these actual frequencies are called the **observed frequencies**.

- ✓ The frequencies obtained by calculation (as if there were no preference) are called the **expected frequencies**.

Example 1

To calculate the expected frequencies, there are two rules to follow:

1. If all the expected frequencies are equal, the expected frequency E can be calculated by using $E = n / k$, where n is the total number of observations and k is the number of categories.

1. If all the expected frequencies are not equal, then the expected frequency E can be calculated by $E = n \cdot p$, where n is the total number of observations and p is the probability for that category.

Looking at the number of arrests example, if there were no difference, you would expect 160 ÷ 4 = 40 arrests for each category. That is, approximately 40 people would be arrested for each type of crime. A completed table is shown.

Example1

	Larceny thefts	Property crimes	Drug use	Driving under the influence
Observed	38	50	28	44
Expected	40	40	40	40

The observed frequencies will almost always differ from the expected frequencies due to sampling error; that is, the values differ from sample to sample. But the question is:

- ✓ Are these differences significant (there is a difference in the number of arrests for these types of crimes) or are they due to chance?
- ✓ The Chi-Square goodness-of-fit test will enable the researcher to determine the answer.

Example 1

Before computing the test value, you must state the hypotheses. The null hypothesis should be a statement indicating that there is no difference or no change. For this example, the hypotheses are as follows:

- ✓ *H0*: There is no difference in the number of arrests for each type of crime.
- ✓ *H1*: There is a difference in the number of arrests for each type of crime.

Next, we need a measure of discrepancy between the observed values *O* and the expected values *E*, so we use the test statistic for the Chi-Square goodness-of-fit test.

Formula for the Chi-Square Goodness-of-Fit Test

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

with degrees of freedom equal to the number of categories minus 1, and where

O = observed frequency

E = expected frequency

Example 1

- When there is perfect agreement between the observed and the expected values, $X^2 = 0$. Also, X^2 can never be negative.
- Finally, the test is right-tailed because "*H0*: Good fit" and "*H1*: Not a good fit" mean that X^2 will be small in the first case and large in the second case.
- In the goodness-of-fit test, the degrees of freedom are equal to the number of categories minus 1. In this example, there are four categories; hence, the degrees of freedom are $4 - 1 = 3$. **This is so because the number of subjects in each of the first three categories is free to vary. But in order for the sum to be 160, the total number of subjects in the last category is fixed**. Two assumptions are needed for the goodness-of-fit test. These assumptions are given next.

Assumptions for the Chi-Square Goodness-of-Fit Test

1. The data are obtained from a random sample.
2. The expected frequency for each category must be 5 or more.

The Chi-Square Goodness-of-Fit Test

Procedure Table

The Chi-Square Goodness-of-Fit Test

Step 1 State the hypotheses and identify the claim.

Step 2 Find the critical value from Table G. The test is always right-tailed.

Step 3 Compute the test value.

Find the sum of the $\frac{(O-E)^2}{E}$ values.

Step 4 Make the decision.

Step 5 Summarize the results.

Example – Arrest for Crimes

Is there enough evidence to reject the claim that the number of arrests for each category of crimes is the same? Use $\alpha = 0.05$.

SOLUTION

Step 1 State the hypotheses and identify the claim.

H_0: There is no difference in the number of arrests for each type of crime. (claim)
H_1: There is a difference in the number of arrests for each type of crime.

Step 2 Find the critical value. The degrees of freedom are $4 - 1 = 3$, and at $\alpha = 0.05$ the critical value from Table G in Appendix A is 7.815.

Step 3 Compute the test value. Note the expected values are found by $E = n/k = 160/4 = 40$.

The table looks like this.

	Larceny thefts	Property crimes	Drug use	Driving under the influence
Observed	38	50	28	44
Expected	40	40	40	40

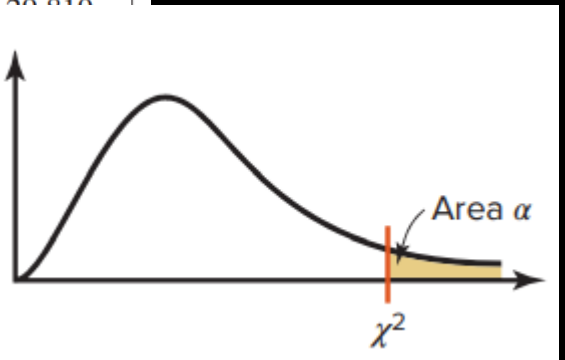
The test value is computed by subtracting the expected value corresponding to the observed value, squaring the result, and dividing by the expected value. Then find the sum of these values.

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(38 - 40)^2}{40} + \frac{(50 - 40)^2}{40} + \frac{(28 - 40)^2}{40} + \frac{(44 - 40)^2}{40}$$
$$= 0.1 + 2.5 + 3.6 + 0.4$$
$$= 6.6$$

The Chi-Square Goodness-of-Fit Test

TABLE G The Chi-Square Distribution

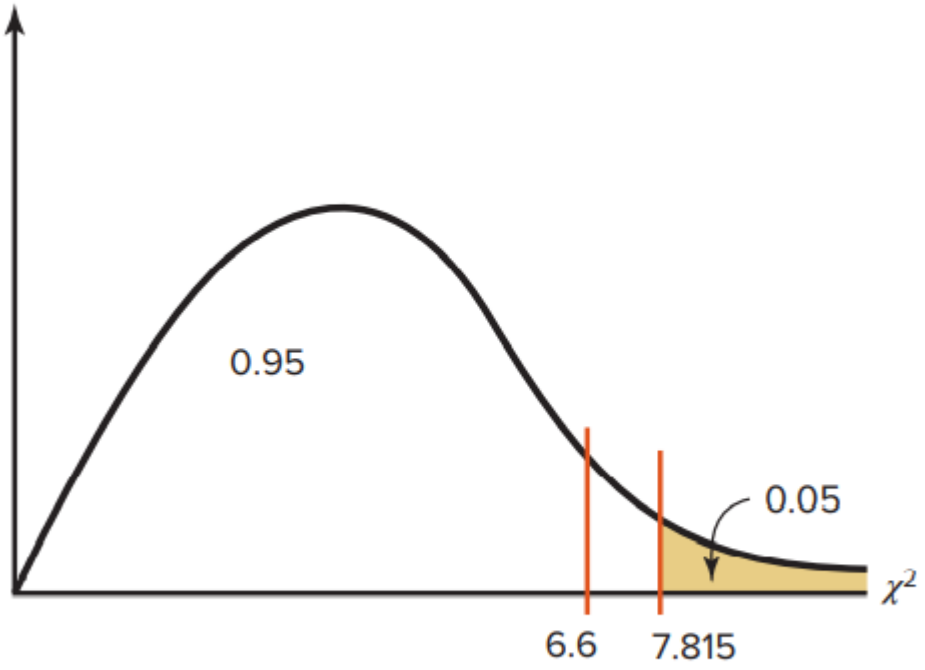
Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	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	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	



Example – Arrest for Crimes

Step 4 Make the decision. The decision is to not reject the null hypothesis since $6.6 < 7.815$, as shown in Figure 11–1.

FIGURE 11–1
Critical and Test Values for Example 11–1

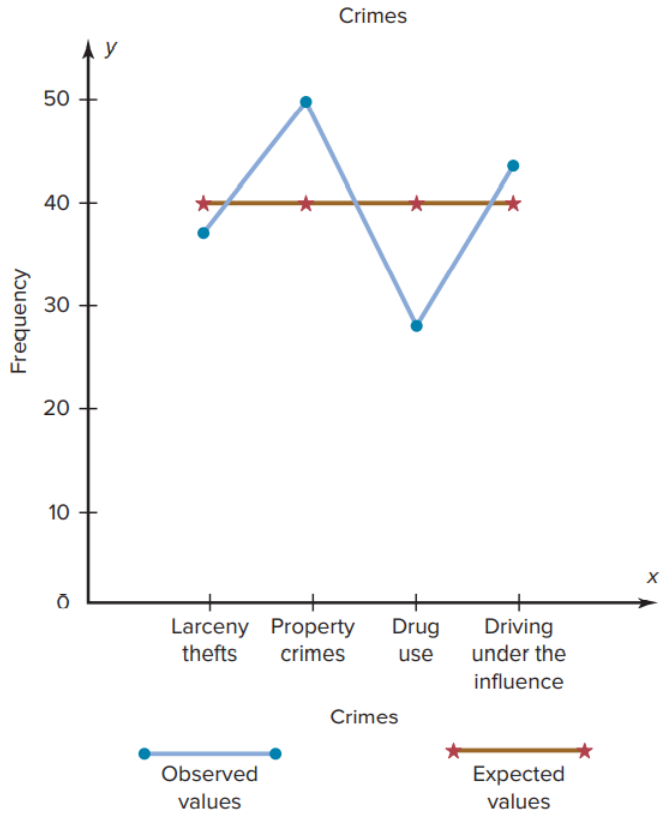


Step 5 Summarize the results. There is not enough evidence to reject the claim that there is no difference in the number of arrests for each type of crime.

P-Value

Also, the *P-value* can be found for this test. In Example 11–1, the test value was 6.6. If you look across the row with *d.f.* = 3 of Table *G*, you will find that 6.6 is between 6.251 and 7.815, that is, between 0.10 and 0.05 at the top of the table.
This corresponds to $0.05 < p < 0.10$. Since the *P-value* is greater than 0.05, the decision is to not reject the null hypothesis.

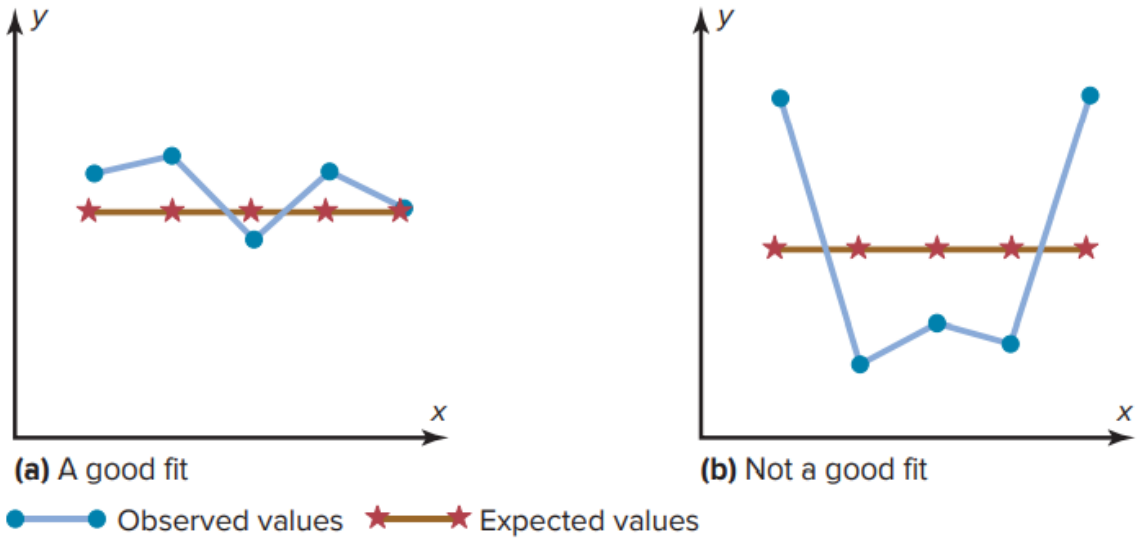
FIGURE 11–2
Graphs of the Observed and Expected Values for Arrests



P-Value

When the observed values and expected values are close together, the Chi-Square test value will be small. Then the decision will be to not reject the null hypothesis—hence, this is "a good fit." See Figure 11–3(a). When the observed values and the expected values are far apart, the Chi-Square test value will be large. Then the null hypothesis will be rejected—hence, there is "not a good fit."

FIGURE 11–3
Results of the Goodness-of-Fit Test



Example 2 – Education Level of Adults

The Census Bureau of the U.S. government found that 13% of adults did not finish high school, 30% graduated from high school only, 29% had some college education but did not obtain a bachelor's degree, and 28% were college graduates.

To see if these proportions were consistent with those people who lived in the Lincoln County area, a local researcher selected a random sample of 300 adults and found that 43 did not finish high school, 76 were high school graduates only, 96 had some college education, and 85 were college graduates.

At $\alpha = 0.10$, test the claim that the proportions are the same for the adults in Lincoln County as those stated by the Census Bureau.

Example2 – Education Level of Adults

SOLUTION

Step 1 State the hypotheses and identify the claim.

H_0: The proportion of people in each category is as follows: 13% did not finish high school, 30% were high school graduates only, 29% had some college education but did not graduate, and 28% had a college degree (claim).

H_1: The distribution is not the same as stated in the null hypothesis.

Step 2 Find the critical value. Since $\alpha = 0.10$ and the degrees of freedom are $4 - 1 = 3$, the critical value is 6.251.

Step 3 Compute the test value. First, we must calculate the expected values. Multiply the total number of people surveyed (300) by the percentages of people in each category.

$$0.13 \times 300 = 39$$
$$0.30 \times 300 = 90$$
$$0.29 \times 300 = 87$$
$$0.28 \times 300 = 84$$

The table looks like this:

Frequency	Did not finish high school	H.S. graduate	Some college	College graduate
Observed	43	76	96	85
Expected	39	90	87	84

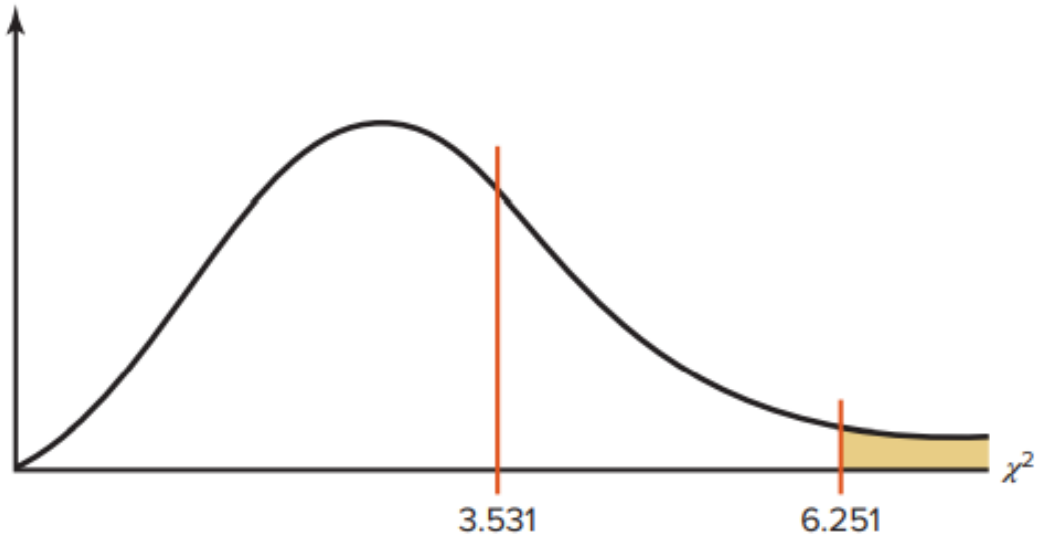
Next calculate the test value.

$$x^2 = \sum \frac{(O-E)^2}{E} = \frac{(43-39)^2}{39} + \frac{(76-90)^2}{90} + \frac{(96-87)^2}{87} + \frac{(85-84)^2}{84}$$
$$= 0.410 + 2.178 + 0.931 + 0.012 = 3.531$$

Example2 – Education Level of Adults

Step 4 Make the decision. Since $3.531 < 6.251$, the decision is not to reject the null hypothesis. See Figure 11–4.

FIGURE 11–4
Critical and Test Values for Example 11–2



Step 5 Summarize the results. There is not enough evidence to reject the claim. It can be concluded that the percentages are not significantly different from those given in the null hypothesis. That is, the proportions are not significantly different from those stated by the U.S. Census Bureau.

Example 3 – Firearm Deaths

A researcher read that firearm-related deaths for people aged 1 to 18 years were distributed as follows: 74% were accidental, 16% were homicides, and 10% were suicides. In her district, there were 68 accidental deaths, 27 homicides, and 5 suicides during the past year. At $\alpha = 0.10$, test the claim that the percentages are equal.

SOLUTION

Step 1 State the hypotheses and identify the claim:

H_0: The deaths due to firearms for people aged 1 through 18 years are distributed as follows: 74% accidental, 16% homicides, and 10% suicides (claim).

H_1: The distribution is not the same as stated in the null hypothesis.

Step 2 Find the critical value. Since $\alpha = 0.10$ and the degrees of freedom are $3 - 1 = 2$, the critical value is 4.605.

Step 3 Compute the test value.

First calculate the expected values, using the formula $E = n \cdot p$ as shown.

$$100 \times 0.74 = 74$$
$$100 \times 0.16 = 16$$
$$100 \times 0.10 = 10$$

The table looks like this.

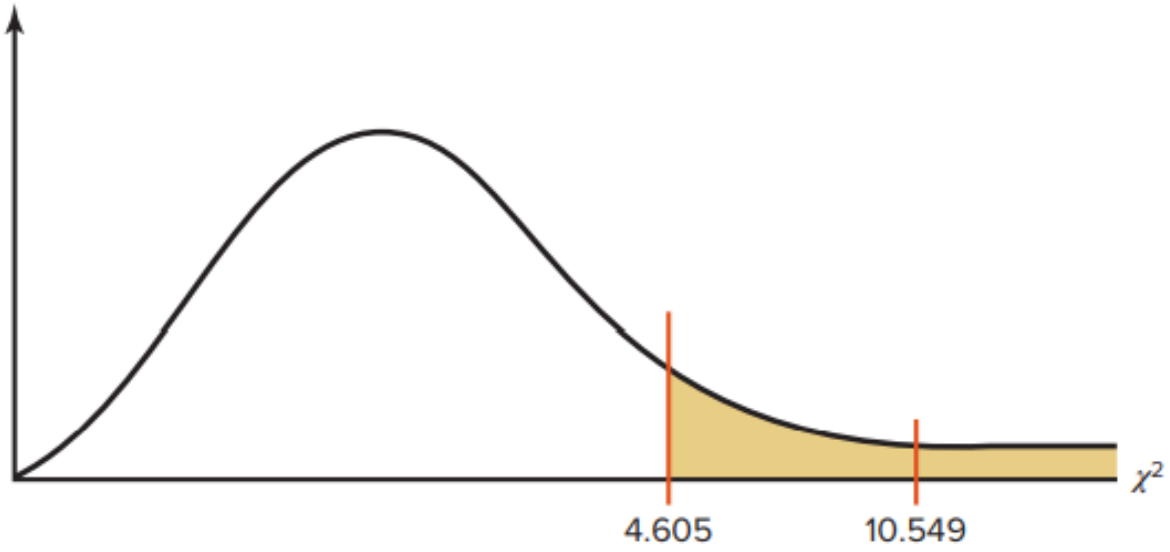
Frequency	Accidental	Homicides	Suicides
Observed	68	27	5
Expected	74	16	10

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$
$$= \frac{(68 - 74)^2}{74} + \frac{(27 - 16)^2}{16} + \frac{(5 - 10)^2}{10}$$
$$= 10.549$$

Example3 – Firearm Deaths

Step 4 Reject the null hypothesis, since 10.549 > 4.605, as shown in Figure 11–5.

FIGURE 11–5
Critical and Test Values for Example 11–3



Step 5 Summarize the results. There is enough evidence to reject the claim that the distribution is 74% accidental, 16% homicides, and 10% suicides.

Test of Normality

The Chi-Square goodness-of-fit test can be used to test a variable to see if it is normally distributed. The null hypotheses are:

- ✓ H0: The variable is normally distributed.
- ✓ H1: The variable is not normally distributed.

- ✓ The procedure is somewhat complicated. It involves finding the expected frequencies for each class of a frequency distribution by using the standard normal distribution. Then the actual frequencies (i.e., observed frequencies) are compared to the expected frequencies, using the Chi-Square goodness-of-fit test.
- ✓ If the observed frequencies are close in value to the expected frequencies, then the Chi-Square test value will be small and the null hypothesis cannot be rejected. In this case, it can be concluded that the variable is approximately normally distributed.
- ✓ On the other hand, if there is a large difference between the observed frequencies and the expected frequencies, then the Chi-Square test value will be larger and the null hypothesis can be rejected. In this case, it can be concluded that the variable is not normally distributed.

Example 4 – Test of Normality

Use Chi-Square to determine if the variable shown in the frequency distribution is normally distributed. Use $\alpha = 0.05$.

Boundaries	Frequency
89.5–104.5	24
104.5–119.5	62
119.5–134.5	72
134.5–149.5	26
149.5–164.5	12
164.5–179.5	4
	Total = 200

SOLUTION

H_0: The variable is normally distributed.
H_1: The variable is not normally distributed.

First find the mean and standard deviation of the variable. (*Note*: s is used to approximate σ.)

Boundaries	f	X_m	$f \cdot X_m$	$f \cdot X_m^2$
89.5–104.5	24	97	2,328	225,816
104.5–119.5	62	112	6,944	777,728
119.5–134.5	72	127	9,144	1,161,288
134.5–149.5	26	142	3,692	524,264
149.5–164.5	12	157	1,884	295,788
164.5–179.5	4	172	688	118,336
	200		24,680	3,103,220

$$\overline{X} = \frac{24{,}680}{200} = 123.4$$

$$s = \sqrt{\frac{200(3{,}103{,}220) - 24{,}680^2}{200(199)}} = \sqrt{290} = 17.03$$

Example 4 – Test of Normality

TABLE E The Standard Normal Distribution

Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

For z values less than −3.49, use 0.0001.

Example4 – Test of Normality

Next find the area under the standard normal distribution, using z values and Table E for each class.

The z score for $x = 104.5$ is found as

$$z = \frac{104.5 - 123.4}{17.03} = -1.11$$

The area for $z < -1.11$ is 0.1335.

The z score for 119.5 is found as

$$z = \frac{119.5 - 123.4}{17.03} = -0.23$$

The area for $-1.11 < z < -0.23$ is $0.4090 - 0.1335 = 0.2755$.

The z score for 134.5 is found as

$$z = \frac{134.5 - 123.4}{17.03} = 0.65$$

The area for $-0.23 < z < 0.65$ is $0.7422 - 0.4090 = 0.3332$.

The z score for 149.5 is found as

$$z = \frac{149.5 - 123.4}{17.03} = 1.53$$

The area for $0.65 < z < 1.53$ is $0.9370 - 0.7422 = 0.1948$.

The z score for 164.5 is found as

$$z = \frac{164.5 - 123.4}{17.03} = 2.41$$

The area for $1.53 < z < 2.41$ is $0.9920 - 0.9370 = 0.0550$.

The area for $z > 2.41$ is $1.0000 - 0.9920 = 0.0080$.

Example 4 – Test of Normality

The critical value in this test has the degrees of freedom equal to the number of categories minus 3 since 1 degree of freedom is lost for each parameter that is estimated. In this case, the mean and standard deviation have been estimated, so 2 additional degrees of freedom are needed. The C.V. with d.f. = 5 − 3 = 2 and $\alpha = 0.05$ is 5.991, so the null hypothesis is rejected. Hence, the distribution can be considered not normally distributed. Note: At $\alpha = 0.01$, the C.V. = 9.210, and the null hypothesis would not be rejected. Hence, we could consider that the variable is normally distributed at $\alpha = 0.01$. **So it is important to decide which level of significance you want to use prior to conducting the test.**

Find the expected frequencies for each class by multiplying the area by 200. The expected frequencies are found by

$$0.1335 \cdot 200 = 26.7$$
$$0.2755 \cdot 200 = 55.1$$
$$0.3332 \cdot 200 = 66.64$$
$$0.1948 \cdot 200 = 38.96$$
$$0.0550 \cdot 200 = 11.0$$
$$0.0080 \cdot 200 = 1.6$$

Note: Since the expected frequency for the last category is less than 5, it can be combined with the previous category.

The table looks like this.

O	24	62	72	26	16
E	26.7	55.1	66.64	38.96	12.6

Finally, find the chi-square test value using the formula $\chi^2 = \sum \frac{(O-E)^2}{E}$.

$$\chi^2 = \frac{(24-26.7)^2}{26.7} + \frac{(62-55.1)^2}{55.1} + \frac{(72-66.64)^2}{66.64} + \frac{(26-38.96)^2}{38.96} + \frac{(16-12.6)^2}{12.6}$$
$$= 6.797$$

Tests Using Contingency Tables

- When data can be tabulated in table form in terms of frequencies, several types of hypotheses can be tested by using the Chi-Square test.
- **Two such tests are the independence of variables test and the homogeneity of proportions test.**
- The test of independence of variables is used to determine whether two variables are independent of or related to each other when a single sample is selected.
- The test of homogeneity of proportions is used to determine whether the proportions for a variable are equal when several samples are selected from different populations.
- Both tests use the Chi-Square distribution and a contingency table, and the test value is found in the same way. The independence test will be explained first.

Tests Using Contingency Tables

The chi-square **independence test** is used to test whether two variables are independent of each other.

Formula for the Chi-Square Independence Test

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

with degrees of freedom equal to the (number of rows minus 1)(number of columns minus 1) and where

O = observed frequency
E = expected frequency

Assumptions for Chi-Square Independence Test

1. The data are obtained from a random sample.
2. The expected value in each cell must be 5 or more. If the expected values are not 5 or more, combine categories.

Tests Using Contingency Tables

The data for the two variables are placed in a **contingency table**. One variable is called the **row variable**, and the other variable is called the **column variable**. The table is called an R × C table, where R is the number of rows and C is the number of columns. For example, a 2 × 3 contingency table would look like this.

Each value in the table is called a cell value. For example, the cell value C2,3 means that it is in the second row (2) and third column (3).

	Column 1	Column 2	Column 3
Row 1	C1,1	C1,2	C1,3
Row 2	C2,1	C2,2	C2,3

Tests Using Contingency Tables

The formula for computing the expected values for each cell is

$$\text{Expected value} = \frac{(\text{row sum})(\text{column sum})}{\text{grand total}}$$

- ✓ The observed values are obtained from the sample data (That is, they are given in the problem.) The expected values are computed from the observed values, and they are based on the assumption that the two variables are independent.
- ✓ As with the Chi-Square goodness-of-fit test, if there is little difference between the observed values and the expected values, then the value of the test statistic will be small and the null hypothesis will not be rejected. Hence, the variables are independent of each other.
- ✓ However, if there are large differences in the observed values and the expected values, then the test statistic will be large and the null hypothesis will be rejected. In this case, there is enough evidence to say that the variables are dependent on or related to each other. **This test is always right-tailed**.

Example 1- Hospitals and Infections

A researcher wishes to see if there is a relationship between the hospital and the number of patient infections. A random sample of 3 hospitals was selected, and the number of infections for a specific year has been reported. The data are shown next.

Hospital	Surgical site infections	Pneumonia infections	Bloodstream infections	Total
A	41	27	51	119
B	36	3	40	79
C	169	106	109	384
Total	246	136	200	582

Source: Pennsylvania Health Care Cost Containment Council.

At $\alpha = 0.05$, can it be concluded that the number of infections is related to the hospital where they occurred?

Example1- Hospitals and Infections

SOLUTION

Step 1 State the hypothesis and identify the claim.

H_0: The number of infections is independent of the hospital.
H_1: The number of infections is dependent on the hospital (claim).

Step 2 Find the critical value. The critical value using Table G at $\alpha = 0.05$ with $(3 - 1)(3 - 1) = (2)(2) = 4$ degrees of freedom is 9.488.

Step 3 Compute the test value. First find the expected values.

$$E_{1,1} = \frac{(119)(246)}{582} = 50.30 \qquad E_{1,2} = \frac{(119)(136)}{582} = 27.81 \qquad E_{1,3} = \frac{(119)(200)}{582} = 40.89$$

$$E_{2,1} = \frac{(79)(246)}{582} = 33.39 \qquad E_{2,2} = \frac{(79)(136)}{582} = 18.46 \qquad E_{2,3} = \frac{(79)(200)}{582} = 27.15$$

$$E_{3,1} = \frac{(384)(246)}{582} = 162.31 \qquad E_{3,2} = \frac{(384)(136)}{582} = 89.73 \qquad E_{3,3} = \frac{(384)(200)}{582} = 131.96$$

The completed table is shown.

Example1- Hospitals and Infections

Hospital	Surgical site infections	Pneumonia infections	Bloodstream infections	Total
A	41 (50.30)	27 (27.81)	51 (40.89)	119
B	36 (33.39)	3 (18.46)	40 (27.15)	79
C	169 (162.31)	106 (89.73)	109 (131.96)	384
Total	246	136	200	582

Then substitute in the formula and evaluate to find the test statistic value.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(41-50.30)^2}{50.30} + \frac{(27-27.81)^2}{27.81} + \frac{(51-40.89)^2}{40.89}$$

$$+ \frac{(36-33.39)^2}{33.39} + \frac{(3-18.46)^2}{18.46} + \frac{(40-27.15)^2}{27.15}$$

$$+ \frac{(169-162.31)^2}{162.31} + \frac{(106-89.73)^2}{89.73} + \frac{(109-131.96)^2}{131.96}$$

$$= 1.719 + 0.024 + 2.500 + 0.204 + 12.948 + 6.082$$

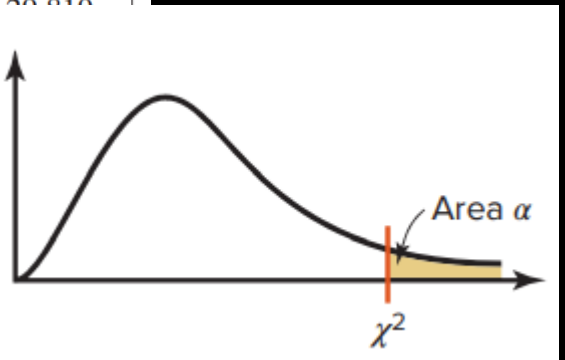
$$+ 0.276 + 2.950 + 3.995$$

$$= 30.698$$

The Chi-Square Goodness-of-Fit Test

TABLE G The Chi-Square Distribution

Degrees of freedom	α									
	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	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	



Example1- Hospitals and Infections

Step 4 Make the decision. The decision is to reject the null hypothesis since $30.698 > 9.488$. That is, the test value lies in the critical region, as shown in Figure 11–7.

FIGURE 11–7
Critical and Test Values for Example 11–5

Step 5 Summarize the results. There is enough evidence to support the claim that the number of infections is related to the hospital where they occurred.

Example 2 – Mental Illness

A researcher wishes to see if there is a difference between men and women in the number of cases of mental disorders. She selects a sample of 30 males and 24 females and classifies them according to their mental disorders. The results are shown.

At $\alpha = 0.10$, can the researcher conclude that there is a difference in the types of disorders based on the gender?

Gender	Anxiety	Depression	Schizophrenia	Total
Male	8	12	10	30
Female	12	9	3	24
Total	20	21	13	54

Example 2-Mental Illness

SOLUTION

Step 1 State the hypotheses and identify the claim.

H_0: The type of mental disorder is independent of the gender of the person.

H_1: The type of mental disorder is related to the gender of the person (claim).

Step 2 Find the critical value. The critical value is 4.605 since the degrees of freedom are $(2-1)(3-1)=2$.

Step 3 Compute the test value. First compute the expected values.

$$E_{1,1}=\frac{(30)(20)}{54}=11.11 \qquad E_{1,2}=\frac{(30)(21)}{54}=11.67 \qquad E_{1,3}=\frac{(30)(13)}{54}=7.22$$

$$E_{2,1}=\frac{(24)(20)}{54}=8.89 \qquad E_{2,2}=\frac{(24)(21)}{54}=9.33 \qquad E_{2,3}=\frac{(24)(13)}{54}=5.78$$

The completed table is shown.

Gender	Anxiety	Depression	Schizophrenia	Total
Male	8 (11.11)	12 (11.67)	10 (7.22)	30
Female	12 (8.89)	9 (9.33)	3 (5.78)	24
Total	20	21	13	54

Example 2 -Mental Illness

The test value is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

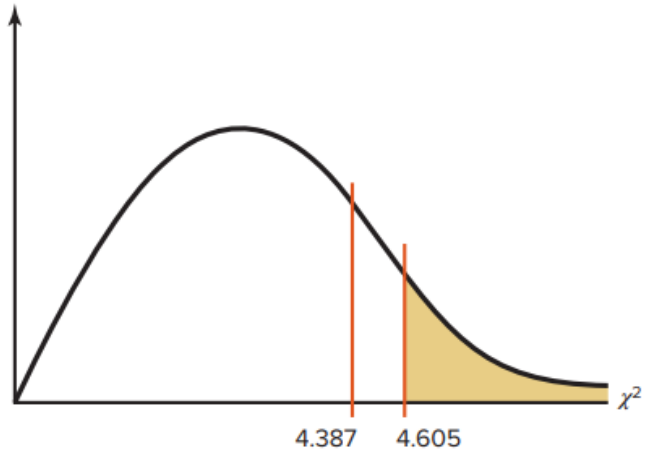
$$= \frac{(8-11.11)^2}{11.11} + \frac{(12-11.67)^2}{11.67} + \frac{(10-7.22)^2}{7.22}$$

$$+ \frac{(12-8.89)^2}{8.89} + \frac{(9-9.33)^2}{9.33} + \frac{(3-5.78)^2}{5.78}$$

$$= 0.871 + 0.009 + 1.070 + 1.088 + 0.012 + 1.337 = 4.387$$

Step 4 Make the decision. The decision is to not reject the null hypothesis since $4.387 < 4.605$. See Figure 11–8.

FIGURE 11–8
Critical and Test Values for Example 11–6



Step 5 Summarize the results. There is not enough evidence to support the claim that the mental disorder is related to the gender of the individual.

In this case, the *P*-value is between 0.10 and 0.90. The TI-84 gives a *P*-value of 0.112. Again, this supports the decision and summary as stated previously.

Test for Homogeneity of Proportions

- ✓ The second Chi-Square test that uses a contingency table is called the **homogeneity of proportions test**.
- ✓ The test of homogeneity of proportions is used to test the claim that different populations have the same proportion of subjects who have a certain attitude or characteristic.
- ✓ If the researcher does not reject the null hypothesis, it can be assumed that the proportions are equal and the differences in them are due to chance. Hence, the proportion of students who smoke is the same for grade levels freshmen through senior. When the null hypothesis is rejected, it can be assumed that the proportions are not all equal.
- ✓ The assumptions for the test of homogeneity of proportions are the same as the assumptions for the Chi-Square test of independence. The procedure for this test is the same as the procedure for the Chi-Square test of independence.

Example 1 – Happiness and Income

A psychologist randomly selected 100 people from each of the four income groups and asked them if they were “very happy.”

For people who made less than $30,000, 24% responded yes.
For people who made $30,000 to $74,999, 33% responded yes.
For people who made $75,000 to $90,999, 38% responded yes,
and for people who made $100,000 or more, 49% responded yes.

At $\alpha = 0.05$, test the claim that there is no difference in the proportion of people in each economic group who were very happy.

Example 1 – Happiness ad Income

SOLUTION

It is necessary to make a table showing the number of people in each group who responded yes and the number of people in each group who responded no.

For group 1, 24% of the people responded yes, so 24% of 100 = 0.24(100) = 24 responded yes and 100 – 24 = 76 responded no.

For group 2, 33% of the people responded yes, so 33% of 100 = 0.33(100) = 33 responded yes and 100 – 33 = 67 responded no.

For group 3, 38% of the people responded yes, so 38% of 100 = 0.38(100) = 38 people responded yes and 100 – 38 = 62 people responded no.

For group 4, 49% of the people responded yes, so 49% of 100 = 0.49(100) = 49 responded yes, and 100 – 49 = 51 people responded no.

Tabulate the data in a table, and find the sums of the rows and columns as shown.

Household income	Less than $30,000	$30,000–$74,999	$75,000–$99,999	$100,000 or more	Total
Yes	24	33	38	49	144
No	76	67	62	51	256
	100	100	100	100	400

Source: Based on information from Princeton Survey Research Associates International.

Example 1 – Happiness ad Income

Step 1 State the hypotheses and identify the claim.

H_0: $p_1 = p_2 = p_3 = p_4$ (claim)

H_1: At least one proportion differs from the others.

Step 2 Find the critical value. The formula for the degrees of freedom is the same as before: $(R - 1)(C - 1) = (2 - 1)(4 - 1) = 1(3) = 3$. The critical value is 7.815.

Step 3 Compute the test value. Since we want to test the claim that the proportions are equal, we use the expected value as $\frac{1}{4} \cdot 400 = 100$. First compute the expected values as shown previously.

$$E_{1,1} = \frac{(144)(100)}{400} = 36 \qquad E_{1,2} = \frac{(144)(100)}{400} = 36 \qquad E_{1,3} = \frac{(144)(100)}{400} = 36 \qquad E_{1,4} = \frac{(144)(100)}{400} = 36$$

$$E_{2,1} = \frac{(256)(100)}{400} = 64 \qquad E_{2,2} = \frac{(256)(100)}{400} = 64 \qquad E_{2,3} = \frac{(256)(100)}{400} = 64 \qquad E_{2,4} = \frac{(256)(100)}{400} = 64$$

Example 1 – Happiness ad Income

The completed table is shown.

Household income	Less than $30,000	$30,000–$74,999	$75,000–$99,999	$100,000 or more	Total
Yes	24 (36)	33 (36)	38 (36)	49 (36)	144
No	76 (64)	67 (64)	62 (64)	51 (64)	256
	100	100	100	100	400

Next calculate the test value.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(24-36)^2}{36} + \frac{(33-36)^2}{36} + \frac{(38-36)^2}{36} + \frac{(49-36)^2}{36}$$

$$+ \frac{(76-64)^2}{64} + \frac{(67-64)^2}{64} + \frac{(62-64)^2}{64} + \frac{(51-64)^2}{64}$$

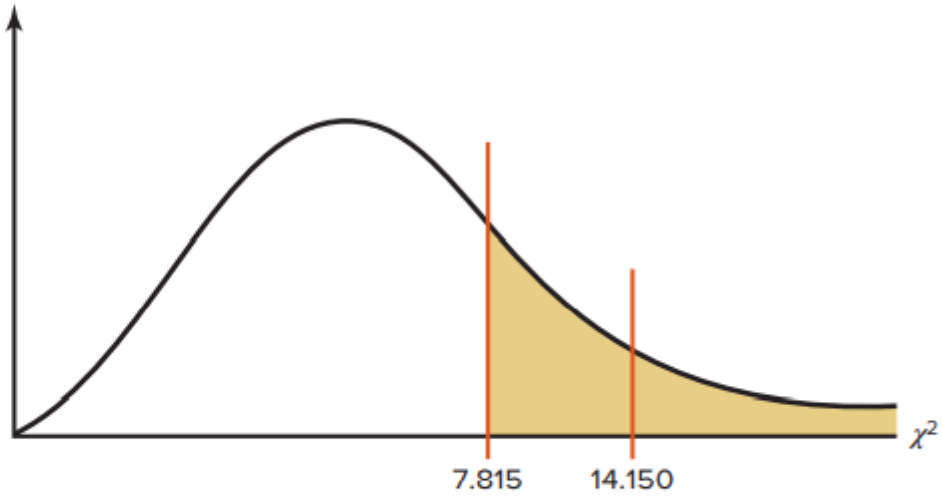
$$= 4.000 + 0.250 + 0.111 + 4.694 + 2.250 + 0.141 + 0.063 + 2.641$$

$$= 14.150$$

Step 4 Make the decision. Reject the null hypothesis since $14.150 > 7.815$. See Figure 11–9.

Example 1 – Happiness ad Income

FIGURE 11–9
Critical and Test Values for Example 11–7



Step 5 Summarize the results. There is enough evidence to reject the claim that there is no difference in the proportions. Hence, the incomes seem to make a difference in the proportions.

Test for Homogeneity of Proportions

When the degrees of freedom for a contingency table are equal to 1—that is, the table is a 2×2 table—some statisticians suggest using the *Yates correction for continuity*. The formula for the test is then

$$\chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

Since the chi-square test is already conservative, most statisticians agree that the Yates correction is not necessary. (See Exercise 33 in Extending the Concepts.)

References

- **Elementary Statistics: A Step-by-Step Approach, Allen Bluman, 10th Edition, McGraw Hill, 2017, ISBN 13: 978-1-259-755330, Chapters 11.**