Nonparametric Methods: Goodness-of-Fit Tests

Chapter 15

McGraw-Hill/Irwin

LEARNING OBJECTIVES

LO 15-1 Conduct a test of hypothesis comparing an observed set of frequencies to an expected distribution.

LO 15-2 List and explain the characteristics of the *chi-square distribution*.

LO 15-3 Compute a goodness-of-fit test for unequal expected frequencies.

LO 15-4 Conduct a test of hypothesis to verify that data grouped into a frequency distribution are a sample from a normal distribution.

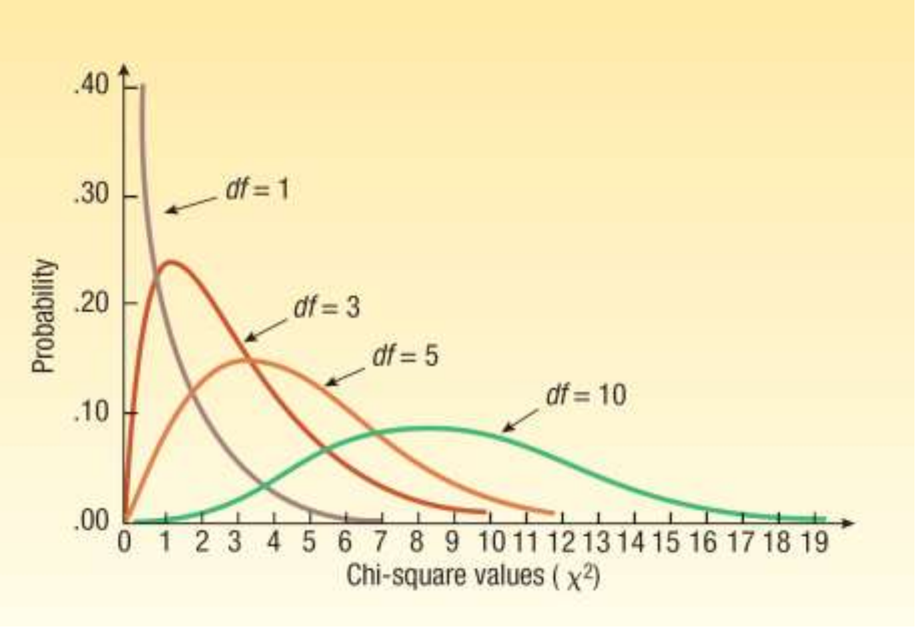
LO 15-5 Use graphical and statistical methods to determine whether a set of sample data is from a normal distribution.

LO 15-6 Perform a chi-square test for independence on a contingency table.

Characteristics of the Chi-Square Distribution

The major characteristics of the chi-square distribution:

- Positively skewed.
- Non-negative.
- Based on degrees of freedom.
- When the degrees of freedom change, a new distribution is created.



Goodness-of-Fit Test: Comparing an Observed Set of Frequencies to an Expected Distribution

- Let f_0 and f_e be the observed and expected frequencies, respectively.
- Hypotheses:
  - H_0: There is no difference between the observed and expected frequencies.
  - H_1: There is a difference between the observed and the expected frequencies.

Goodness-of-fit Test: Comparing an Observed Set of Frequencies to an Expected Distribution

The test statistic is:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

The critical value is a chi-square value with (k – 1) degrees of freedom, where k is the number of categories.

LO 15-1

Goodness-of-Fit Example

Bubba, the owner of Bubba's Fish and Pasta, a chain of restaurants located along the Gulf Coast of Florida, is considering adding steak to his menu. Before doing so, he decides to hire Magnolia Research, LLC to conduct a survey of adults about their favorite meal when eating out. Magnolia selected a sample 120 adults and asked each to indicate their favorite meal when dining out. The results are reported in the table.

Favorite Entrée	Frequency
Chicken	32
Fish	24
Meat	35
Pasta	29
Total	120

Is it reasonable to conclude there is no preference among the four entrées?

LO 15-1

Goodness-of-Fit Example

Step 1: State the null hypothesis and the alternate hypothesis.

H_0: There is no difference between f_o and f_e.

H_1: There is a difference between f_o and f_e.

Step 2: Select the level of significance.

$\alpha = 0.05$ as stated in the problem.

Step 3: Select the test statistic.

The test statistic follows the chi-square distribution, designated as χ^2.

$$\chi^2 = \sum\left[\frac{(f_o - f_e)^2}{f_e}\right]$$

LO 15-1

Goodness-of-Fit Example

Step 4: Formulate the decision rule.

Reject H_0 if $\chi^2 > \chi^2_{\alpha,k-1}$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{\alpha,k-1}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{.05,4-1}$$

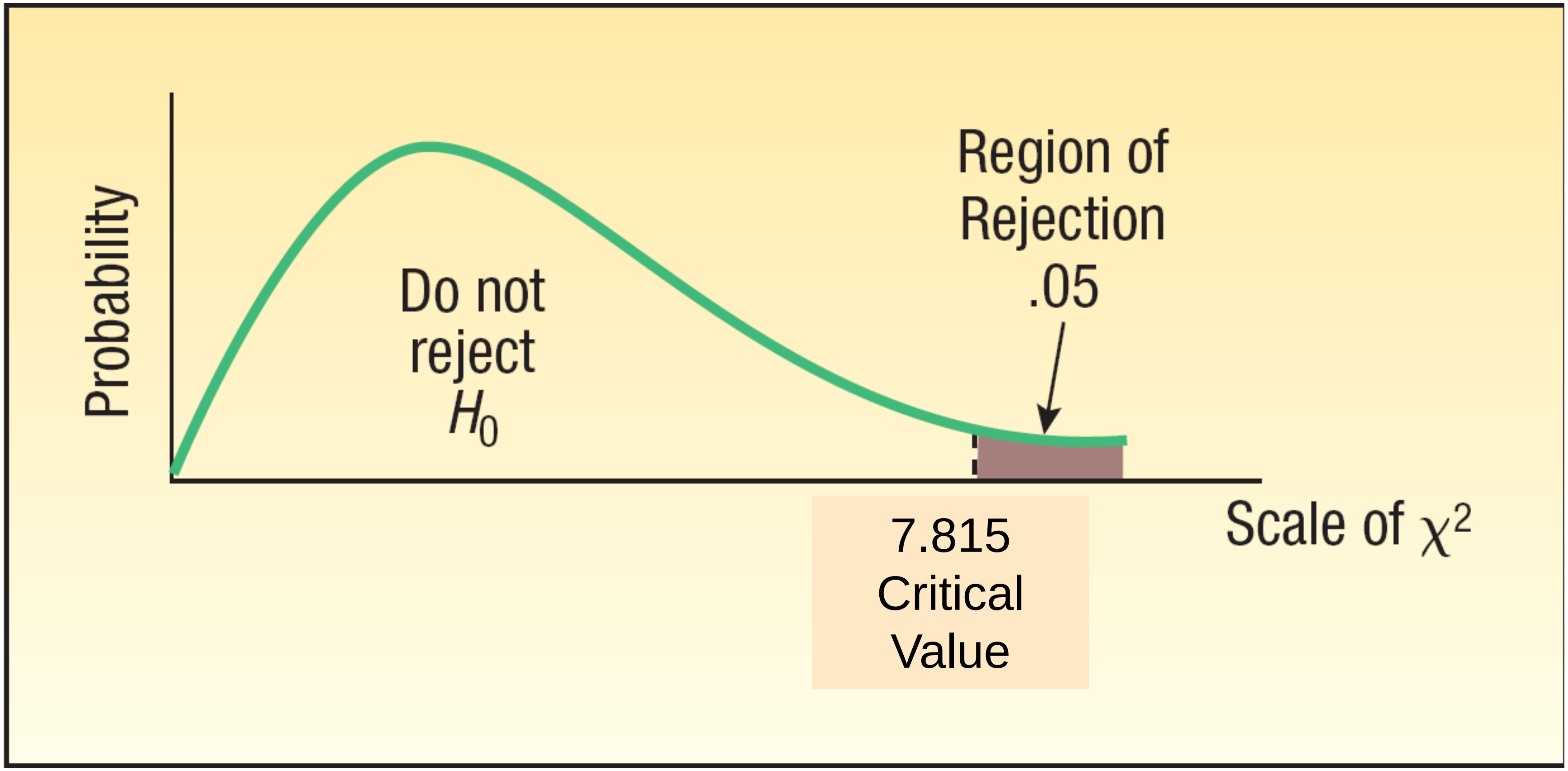
$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{.05,3}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > 7.815$$

A Portion of the Chi-Square Table

Degrees of Freedom df	Right-Tail Area .10	.05	.02	.01
1	2.706	3.841	5.412	6.635
2	4.605	5.991	7.824	9.210
3	6.251	7.815	9.837	11.345
4	7.779	9.488	11.668	13.277

Goodness-of-Fit Example



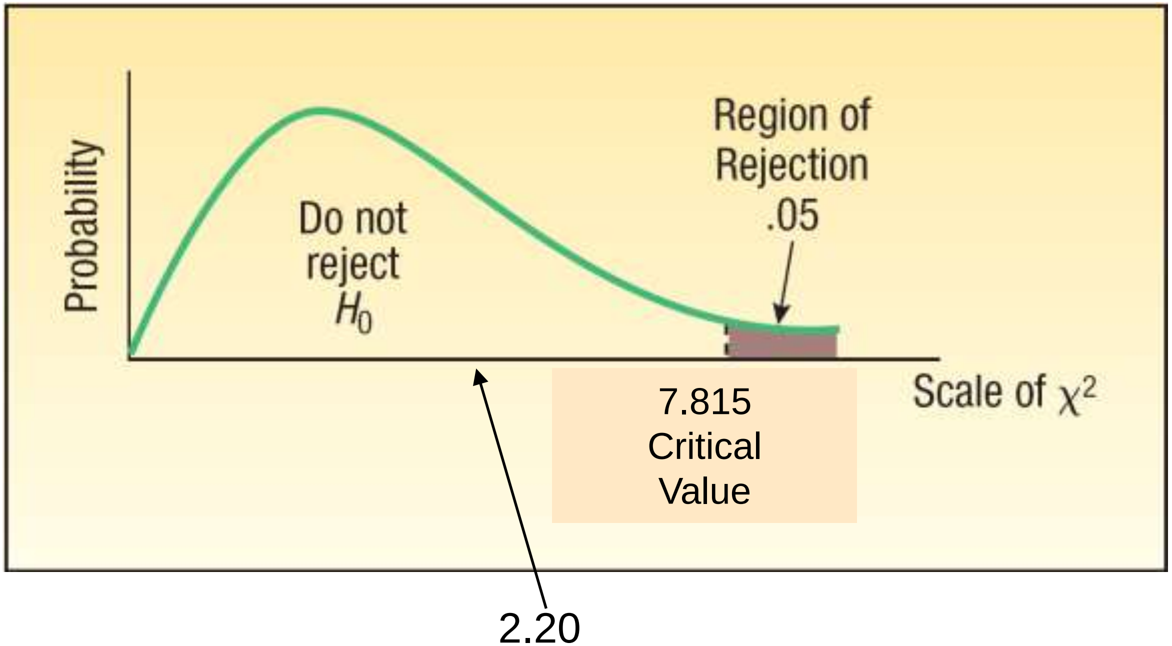
Goodness-of-Fit Example

Step 5: Compute the value of the chi-square statistic and make a decision.

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

Favorite Entrée	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Chicken	32	30	2	4	0.133
Fish	24	30	−6	36	1.200
Meat	35	30	5	25	0.833
Pasta	29	30	−1	1	0.033
Total	120	120	0		2.200

Goodness-of-Fit Example



The computed χ^2 of 2.20 is less than the critical value of 7.815. The decision, therefore, is to fail to reject H_0 at the .05 level.

Conclusion:
The difference between the observed and the expected frequencies is due to chance. There is no difference in preference toward the four entrées.

Chi-square – MegaStat

Goodness-of-Fit Test

observed	expected	O − E	$(O - E)^2/E$	% of chisq
32	30.000	2.000	0.133	6.06
24	30.000	−6.000	1.200	54.55
35	30.000	5.000	0.833	37.88
29	30.000	−1.000	0.033	1.52
120	120.000	0.000	2.200	100.00

2.20	chi-square
3	df
.5319	p-value

Goodness-of-Fit Test: Unequal Expected Frequencies

- Let f_0 and f_e be the observed and expected frequencies, respectively.
- Hypotheses:
  - H_0: There is no difference between the observed and expected frequencies.
  - H_1: There is a difference between the observed and the expected frequencies.

Goodness-of-Fit Test: Unequal Expected Frequencies – Example

The American Hospital Administrators Association (AHAA) reports the following information concerning the number of times senior citizens are admitted to a hospital during a one-year period. Forty percent are not admitted; 30 percent are admitted once; 20 percent are admitted twice, and the remaining 10 percent are admitted three or more times.

A survey of 150 residents of Bartow Estates, a community devoted to active seniors located in central Florida, revealed 55 residents were not admitted during the last year, 50 were admitted to a hospital once, 32 were admitted twice, and the rest of those in the survey were admitted three or more times.

Can we conclude the survey at Bartow Estates is consistent with the information suggested by the AHAA? Use the .05 significance level.

LO 15-3

Goodness-of-Fit Test: Unequal Expected Frequencies – Example

Step 1: State the null hypothesis and the alternate hypothesis.

H_0: There is no difference between local and national experience for hospital admissions.

H_1: There is a difference between local and national experience for hospital admissions.

Step 2: Select the level of significance.

$\alpha = 0.05$ as stated in the problem.

Step 3: Select the test statistic.

The test statistic follows the chi-square distribution, designated as χ^2.

CHI-SQUARE TEST STATISTIC

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right]$$

Goodness-of-Fit Test: Unequal Expected Frequencies – Example

Step 4: Formulate the decision rule.

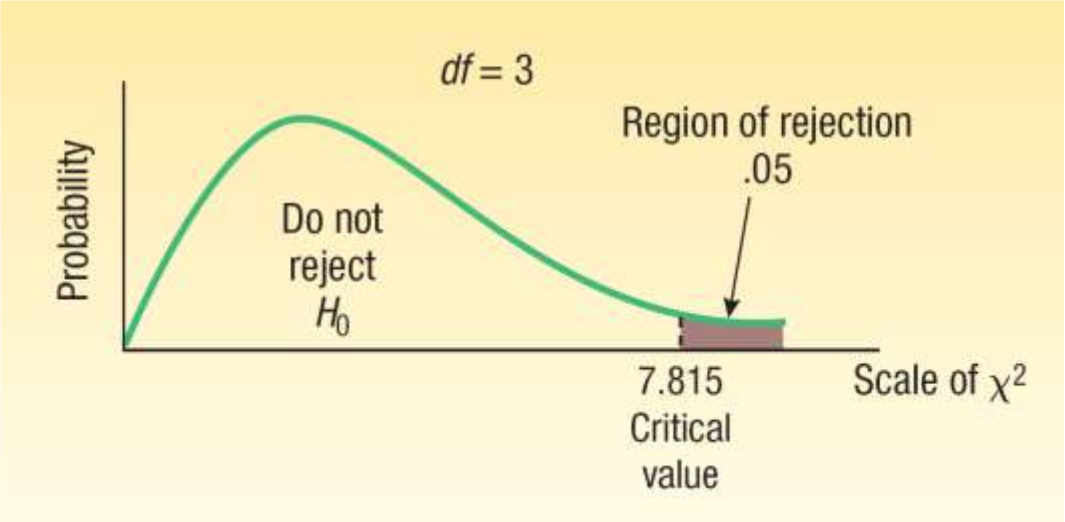
Reject H_0 if $\chi^2 > \chi^2_{\alpha,k-1}$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{\alpha,k-1}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{.05,4-1}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{.05,3}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > 7.815$$



Goodness-of-Fit Test: Unequal Expected Frequencies – Example

Distribution stated in the problem

Frequencies observed in a sample of 150 Bartow residents

Expected frequencies of sample if the distribution stated in the null hypothesis is correct

Number of Times Admitted	AHAA Percent of Total	Number of Bartow Residents (f_o)	Expected Number of Residents (f_e)
0	40	55	60
1	30	50	45
2	20	32	30
3 or more	10	13	15
Total	100	150	150

Computation of f_e
0.40 X 150 = 60
0.30 X 150 = 45
0.30 X 150 = 30
0.10 X 150= 15

Goodness-of-Fit Test: Unequal Expected Frequencies – Example

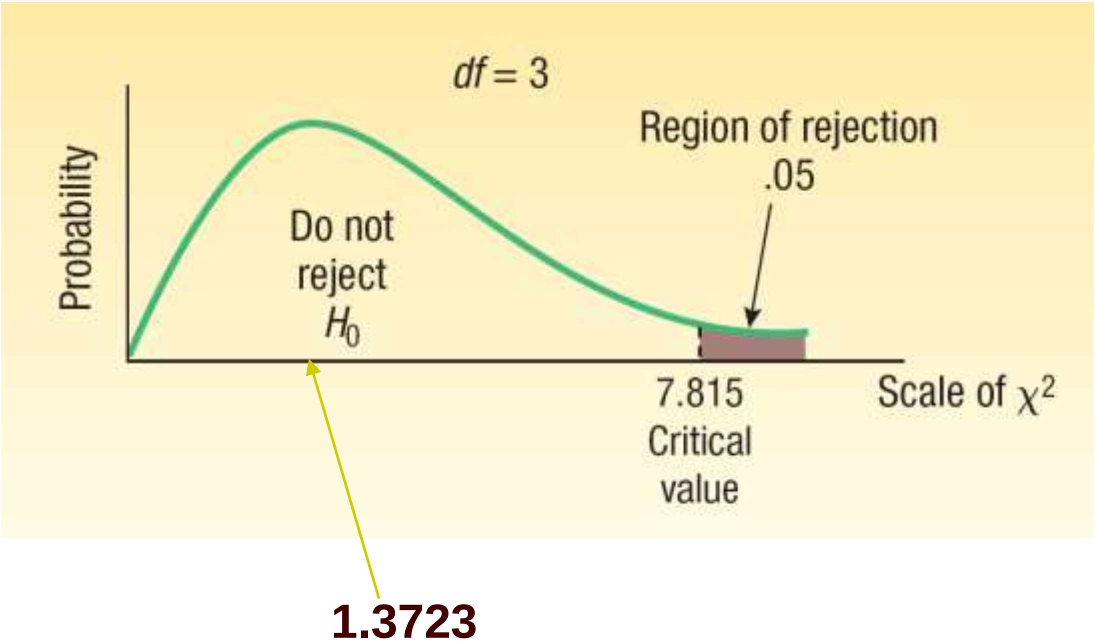
Step 5: Compute the value of the Chi-square statistic and make a decision

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

Number of Times Admitted	(f_o)	(f_e)	$f_o - f_e$	$(f_o - f_e)^2/f_e$
0	55	60	−5	0.4167
1	50	45	5	0.5556
2	32	30	2	0.1333
3 or more	13	15	−2	0.2667
Total	150	150	0	1.3723

Computed χ^2

Goodness-of-Fit Test: Unequal Expected Frequencies – Example



The computed χ^2 of 1.3723 is in the "Do not reject H_0" region. The difference between the observed and the expected frequencies is due to chance.

We conclude that there is no evidence of a difference between the local and national experience for hospital admissions.

Testing the Hypothesis that a Distribution of Data Is from a Normal Population

Recall the frequency distribution of Applewood's profits from the sale of 180 vehicles. The frequency distribution is repeated below.

Is it reasonable to conclude that the profit data is a sample obtained from a normal population?

Profit	Frequency
$ 200 up to $ 600	8
600 up to 1,000	11
1,000 up to 1,400	23
1,400 up to 1,800	38
1,800 up to 2,200	45
2,200 up to 2,600	32
2,600 up to 3,000	19
3,000 up to 3,400	4
Total	180

Testing the Hypothesis that a Distribution of Data Is from a Normal Population

$$z = \frac{X-\mu}{\sigma} = \frac{\$200-\$1843.17}{643.63}$$

$$z = \frac{X-\mu}{\sigma} = \frac{\$600-\$1843.17}{643.63} = -1.93$$

Step 1: Calculate the probabilities for each class.

Convert each class limit into a *z*-score using a mean of $1,843.17 and a standard deviation of $643.63, then find the probability.

Profit	z-Values	Area	Found by	Expected Frequency
Under $200	Under −2.55	.0054	0.5000 − 0.4946	0.97
$ 200 up to $ 600	−2.55 up to −1.93	.0214	0.4946 − 0.4732	3.85
600 up to 1,000	−1.93 up to −1.31	.0683	0.4732 − 0.4049	12.29
1,000 up to 1,400	−1.31 up to −0.69	.1500	0.4049 − 0.2549	27.00
1,400 up to 1,800	−0.69 up to −0.07	.2270	0.2549 − 0.0279	40.86
1,800 up to 2,200	−0.07 up to 0.55	.2367	0.0279 + 0.2088	42.61
2,200 up to 2,600	0.55 up to 1.18	.1722	0.3810 − 0.2088	31.00
2,600 up to 3,000	1.18 up to 1.80	.0831	0.4641 − 0.3810	14.96
3,000 up to 3,400	1.80 up to 2.42	.0281	0.4922 − 0.4641	5.06
3,400 or more	2.42 or more	.0078	0.5000 − 0.4922	1.40
Total		1.0000		180.00

Testing the Hypothesis that a Distribution of Data Is from a Normal Population

Step 2: Use these probabilities to compute the expected frequencies for each class.

0.0214 X 180 = 3.852

Profit	z-Values	Area	Found by	Expected Frequency
Under $200	Under −2.55	.0054	0.5000 − 0.4946	0.97
$ 200 up to $ 600	−2.55 up to −1.93	.0214	0.4946 − 0.4732	3.85
600 up to 1,000	−1.93 up to −1.31	.0683	0.4732 − 0.4049	12.29
1,000 up to 1,400	−1.31 up to −0.69	.1500	0.4049 − 0.2549	27.00
1,400 up to 1,800	−0.69 up to −0.07	.2270	0.2549 − 0.0279	40.86
1,800 up to 2,200	−0.07 up to 0.55	.2367	0.0279 + 0.2088	42.61
2,200 up to 2,600	0.55 up to 1.18	.1722	0.3810 − 0.2088	31.00
2,600 up to 3,000	1.18 up to 1.80	.0831	0.4641 − 0.3810	14.96
3,000 up to 3,400	1.80 up to 2.42	.0281	0.4922 − 0.4641	5.06
3,400 or more	2.42 or more	.0078	0.5000 − 0.4922	1.40
Total		1.0000		180.00

Testing the Hypothesis that a Distribution of Data Is from a Normal Population

Profit	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Under $600	8	4.82	3.18	10.1124	2.098
$ 600 up to $1,000	11	12.29	−1.29	1.6641	.135
1,000 up to 1,400	23	27.00	−4.00	16.0000	.593
1,400 up to 1,800	38	40.86	−2.86	8.1796	.200
1,800 up to 2,200	45	42.61	2.39	5.7121	.134
2,200 up to 2,600	32	31.00	1.00	1.0000	.032
2,600 up to 3,000	19	14.96	4.04	16.3216	1.091
3,000 and over	4	6.46	−2.46	6.0516	.937
Total	180	180.00	0		5.220

Step 3: Compute the chi-square statistic using:

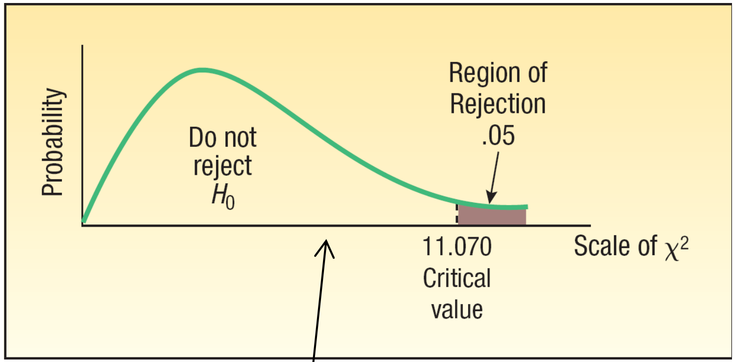
$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(8 - 4.82)^2}{4.82} + \ldots + \frac{(4 - 6.46)^2}{6.46} = 5.220$$

Testing the Hypothesis that a Distribution of Data Is from a Normal Population

Step 4: Compare the computed statistic to the critical statistic and make a statistical conclusion:

H_0: The population of profits follows the normal distribution

H_1: The population of profits does not follow the normal distribution



$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(8-4.82)^2}{4.82} + \ldots + \frac{(4-6.46)^2}{6.46} = 5.220$$

LO 15-5 Use graphical methods to determine whether a set of sample data is from a normal distribution.

Graphical Approach to Confirm Normality: Anderson-Darling Test

Step 1: Create 2 cumulative distributions.

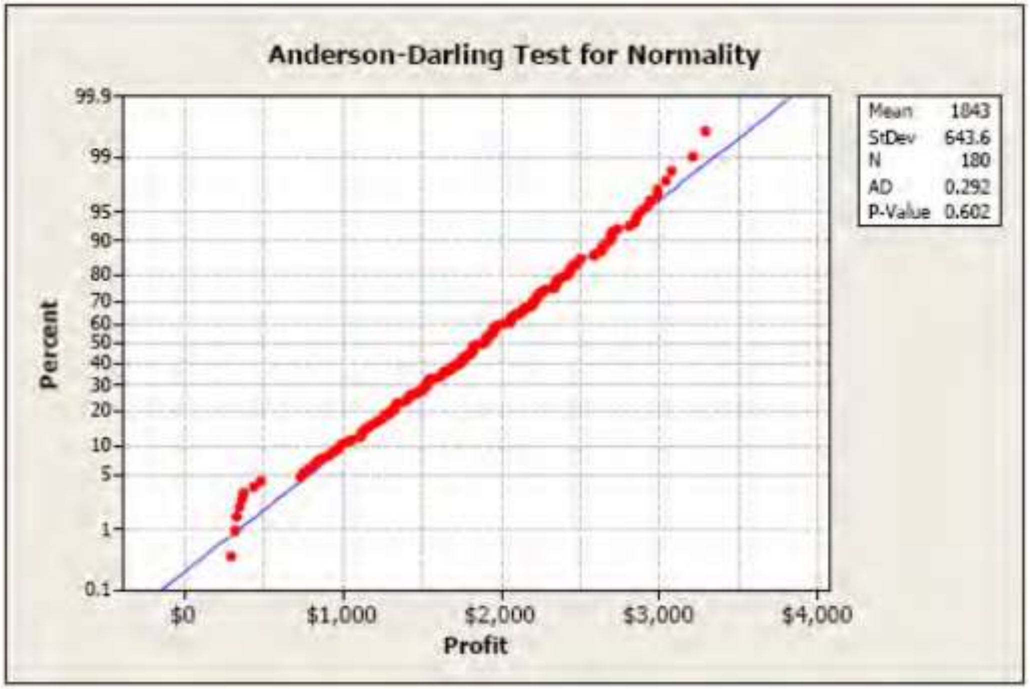
a. Cumulative distribution of the raw data.

b. Cumulative normal distribution.

Step 2: Compare the 2 cumulative distributions.

a. Search the largest absolute numerical difference between the 2 distributions.

b. Using a statistical test, if the difference is large, then we reject the null hypothesis that the data is normally distributed.



The red dots in the graph represent the profit of each of the 180 vehicles from the Applewood Auto Group, and the blue line, which is mostly covered by the red dots, represents a normal cumulative distribution. The graph shows that the profit data closely follows the blue line and that the distribution of profits follows a normal distribution rather closely.

LO 15-6 Perform a chi-square test for independence on a contingency table.

Contingency Table Analysis

A **contingency table** is used to investigate whether two traits or characteristics are related. Each observation is classified according to two criteria. We use the usual hypothesis testing procedure.

- The ***degrees of freedom*** is equal to:

  (number of rows – 1)(number of columns – 1).

- The ***expected frequency*** is computed as:

EXPECTED FREQUENCY

$$f_e = \frac{\text{(Row total)(Column total)}}{\text{Grand total}}$$

Contingency Analysis

We can use the chi-square statistic to formally test for a relationship between two nominal-scaled variables. To put it another way, Is one variable *independent* of the other?

- Ford Motor Company runs an assembly plant in Dearborn, Michigan. The plant operates three shifts per day, 5 days a week. The quality control manager wishes to compare the quality level on the three shifts. Vehicles are classified by quality level (acceptable, unacceptable) and shift (day, afternoon, night). Is there a difference in the quality level on the three shifts? That is, is the quality of the product related to the shift when it was manufactured? Or is the quality of the product independent of the shift on which it was manufactured?

- A sample of 100 drivers, who were stopped for speeding violations, was classified by gender and whether or not the drivers were wearing a seatbelt when stopped. For this sample, is wearing a seatbelt related to gender?

- Does a male released from federal prison make a different adjustment to civilian life if he returns to his hometown or if he goes elsewhere to live? The two variables are adjustment to civilian life and place of residence. Note that both variables are measured on the nominal scale.

Contingency Analysis – Example

The Federal Correction Agency is investigating the question, “Does a male released from federal prison make a different adjustment to civilian life if he returns to his hometown or if he goes elsewhere to live?” To put it another way, is there a relationship between adjustment to civilian life and place of residence after release from prison? Use the .01 significance level.

Contingency Analysis – Example

The agency's psychologists interviewed 200 randomly selected former prisoners. Using a series of questions, the psychologists classified the adjustment of each individual to civilian life as outstanding, good, fair, or unsatisfactory.

The classifications for the 200 former prisoners were tallied as follows. Joseph Camden, for example, returned to his hometown and has shown outstanding adjustment to civilian life. His case is one of the 27 tallies in the upper left box (circled).

Residence after Release from Prison	Adjustment to Civilian Life			
	Outstanding	Good	Fair	Unsatisfactory
Hometown	卌 卌 卌 卌 卌 ||	卌 卌 卌 卌 卌 卌 卌	卌 卌 卌 卌 卌 卌 |||	卌 卌 卌 卌 卌
Not hometown	卌 卌 |||	卌 卌 卌	卌 卌 卌 卌 卌 ||	卌 卌 卌 卌 卌

Residence after Release from Prison	Adjustment to Civilian Life				
	Outstanding	Good	Fair	Unsatisfactory	Total
Hometown	27	35	33	25	120
Not hometown	13	15	27	25	80
Total	40	50	60	50	200

Contingency Analysis – Example

Step 1: State the null hypothesis and the alternate hypothesis.

H_0: There is no relationship between adjustment to civilian life and where the individual lives after being released from prison.

H_1: There is a relationship between adjustment to civilian life and where the individual lives after being released from prison.

Step 2: Select the level of significance.

$\alpha = 0.01$ as stated in the problem.

Step 3: Select the test statistic.

The test statistic follows the chi-square distribution, designated as χ^2.

CHI-SQUARE TEST STATISTIC

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right]$$

Contingency Analysis – Example

Step 4: Formulate the decision rule.

$$\text{Reject } H_0 \text{ if } \chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{\alpha,(2-1)(4-1)}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{.01,(1)(3)}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{.01,3}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > 11.345$$

Computing Expected Frequencies (f_e)

EXPECTED FREQUENCY $$f_e = \frac{\text{(Row total)(Column total)}}{\text{Grand total}}$$

Residence after Release from Prison	Outstanding f_o	Outstanding f_e	Good f_o	Good f_e	Fair f_o	Fair f_e	Unsatisfactory f_o	Unsatisfactory f_e	Total f_o	Total f_e
Hometown	27	24	35	30	33	36	25	30	120	120
Not hometown	13	16	15	20	27	24	25	20	80	80
Total	40	40	50	50	60	60	50	50	200	200

Adjustment to Civilian Life

$\frac{(120)(50)}{200}$ (Hometown, Good f_e = 30)

$\frac{(80)(50)}{200}$ (Not hometown, Good f_e = 20)

Must be equal (Outstanding totals 40 and 40)

Must be equal (Total 200 and 200)

Computing the Chi-square Statistic

Residence after Release from Prison	Adjustment to Civilian Life: Outstanding f_o	Outstanding f_e	Good f_o	Good f_e	Fair f_o	Fair f_e	Unsatisfactory f_o	Unsatisfactory f_e	Total f_o	Total f_e
Hometown	27	24	35	30	33	36	25	30	120	120
Not hometown	13	16	15	20	27	24	25	20	80	80
Total	40	40	50	50	60	60	50	50	200	200

Must be equal (Outstanding totals f_o and f_e)

$\frac{(80)(50)}{200}$ (Not hometown, Good f_e)

Must be equal (Total f_o and f_e)

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right]$$

Starting with the upper left cell:

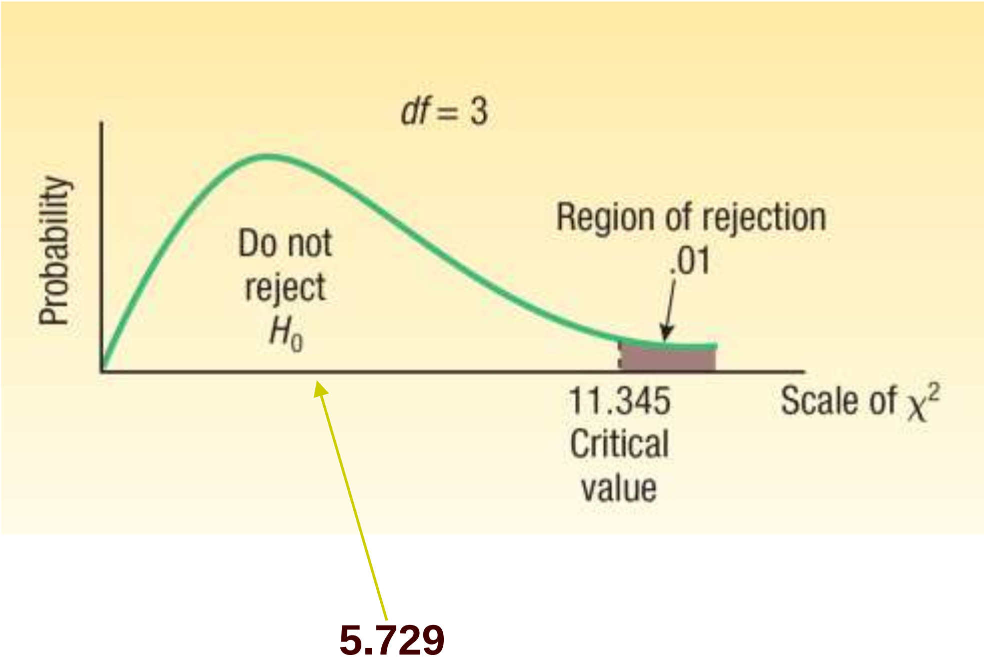
$$\chi^2 = \frac{(27-24)^2}{24} + \frac{(35-30)^2}{30} + \frac{(33-36)^2}{36} + \frac{(25-30)^2}{30}$$

$$+ \frac{(13-16)^2}{16} + \frac{(15-20)^2}{20} + \frac{(27-24)^2}{24} + \frac{(25-20)^2}{20}$$

$$= 0.375 + 0.833 + 0.250 + 0.833 + 0.563 + 1.250 + 0.375 + 1.250$$

$$= 5.729$$

Conclusion



The computed χ^2 of 5.729 is in the "Do not reject H_0" region. The null hypothesis is not rejected at the .01 significance level.

We conclude there is no evidence of a relationship between adjustment to civilian life and where the prisoner resides after being released from prison. For the Federal Correction Agency's advisement program, adjustment to civilian life is not related to where the ex-prisoner lives.

Contingency Analysis – Minitab