CHAPTER **14**

Chi-Square

Chapter Outline

14.1 The Goodness-of-Fit Test

Learning Objectives

- Learn how to use a chi square test to evalute the fit of a hypothesized distribution.
- Identify the conditions which must be satisfied when using the chi-square test.
- Evaluate a hypothesis using the goodness-of-fit test.

Introduction

In previous lessons, we learned that there are several different tests that we can use to analyze data and test hypotheses. The type of test that we choose depends on the data available and what question we are trying to answer. We analyze simple descriptive statistics, such as the mean, median, mode, and standard deviation to give us an idea of the distribution and to remove outliers, if necessary. We calculate probabilities to determine the likelihood of something happening. Finally, we use regression analysis to examine the relationship between two or more continuous variables. But suppose you wanted to evaluate a recent statistic stating that iOS represents 32% and Android 51% of active smart phones. You would like to know if the statistic actually reflects the distribution of phones among your friends. How could you evaluate the data you collect to see if it supports this hypothesis?

Chi Square Statistic

The primary difference between a chi-square test and the tests we have worked with before is that chi square tests are for used for categorical data. The chi-square test can be used to estimate how closely the distribution of a categorical variable matches an expected distribution (the goodness-of-fit test), or to estimate whether two categorical variables are independent of one another (the test of independence). The chi square test of independence is a natural extension of what we did earlier with contingency tables to examine whether or not two variables appeared to be independent of each outher. In this lesson, we will examine the goodness-of-fit test in more detail.

Goodness-of-Fit Test

A goodness of fit test is a test that is concerned with the distribution of one categorical variable. The null and alternative hypotheses reflect this focus:

H_0 : The population distribution of the variable is the same as the proposed distribution

H_A : The distributions are different

The Greek letter "chi", written as χ, is the symbol used to identify a **chi-square statistic**, which we will use here to evaluate how well a set of observed categorical data fits a hypothesized distribution. The Chi-Square statistic is actually pretty straightforward to calculate:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

where

Observed = actual count values in each category
Expected = the predicted (expected) counts in each category if the null hypothesis were true

Conducting a **Chi-Square test** is much like conducting a Z-test or T-test. We will follow the same basic series of steps and compare a calculated value to a value on a distribution table to evaluate the probability of getting the results we have if the null hypothesis is true, just as we did with the Z and T tests. Now, we will use something called the chi-square distribution table. Additionally, we will be evaluating the number of *degrees of freedom*, and choosing values from a chart based on the number. For the goodness-of-fit chi-square test, the degrees of freedom are found by taking the number of levels in our categorical variable and subtracting one.

Assumptions of the Chi-Square test

The assumptions of the chi-square test are the same whether we are using the goodness-of-fit or the test-of-independence. The standard assumptions are:

- Random sample.
- Independent observations for the sample (one observation per subject).
- No expected counts less than five.

Notice that the last two assumptions are concerned with the *expected* counts, not the raw observed counts.

Example A

The *American Pet Products Association* conducted a survey in 2011 and determined that 60% of dog owners have only one dog, 28% have two dogs, and 12% have three or more. Supposing that you have decided to conduct your own survey and have collected the data below, determine whether your data supports the results of the *APPA* study. Use a significance level of 0.05.

Data: Out of 129 dog owners, 73 had one dog and 38 had two dogs.

Solution

- **Step 1:** Clearly state the null and alternative hypotheses.

    H_0: *The proportion of dog owners with one, two or three dogs is 0.60, 0.28 and 0.12 respectively.*

    H_1: *The proportion of dog owners with one, two or three dogs does not match the proposed model.*

- **Step 2:** Identify an appropriate test and significance level.

    A Chi-Square goodness of fit test is appropriate because we are examining the distribution of a single categorical variable. In the absence of a stated significance level in the problem, we assume the default 0.05.

- **Step 3:** Analyze sample data.

    Create a table to organize data and compare the observed data to the expected data:

TABLE 14.1:

	One Dog	Two Dogs	3+ Dogs	TOTAL
Observed	73	38	18	129
Expected				

To identify the expected values, multiply the expected % by the total number observed:

TABLE 14.2:

	One Dog	Two Dogs	3+ Dogs	TOTAL
Observed	73	38	18	129
Expected	$0.60 \times 129 = 77.4$	$0.28 \times 129 = 36.1$	$0.12 \times 129 = 15.5$	129

To calculate our chi-square statistic, we need to sum the squared difference between each observed and expected value divided by the expected value:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

$$\chi^2 = \frac{(73 - 77.4)^2}{77.4} + \frac{(38 - 36.1)^2}{36.1} + \frac{(18 - 15.5)^2}{15.5}$$

$$\chi^2 = \frac{(-4.4)^2}{77.4} + \frac{(1.9)^2}{36.1} + \frac{(2.5)^2}{15.5}$$

$$\chi^2 = \frac{19.36}{77.4} + \frac{3.61}{36.1} + \frac{6.25}{15.5}$$

$$\chi^2 = 0.2501 + 0.1000 + 0.4032$$

$$\chi^2 = 0.7533$$

Now that we have our chi-square statistic, we need to compare it to the chi-square value for the significance level 0.05. We can use a reference table such as the one below. Just as with the T-tests, we will need to know the *degrees of freedom*, which equal the number of observed category values minus one. In this case, there are three category values: one dog, two dogs, and three or more dogs. The degrees for freedom, therefore, are $3 - 1 = 2$.

TABLE 14.3: Right Tail Probability

DF	0.20	0.10	0.05	0.025	0.01
1	1.64	2.71	3.84	5.02	6.64
2	3.22	4.60	5.99	7.38	9.21
3	4.64	6.25	7.82	9.35	11.34
4	5.99	7.78	9.49	11.14	13.28
5	7.29	9.24	11.07	12.83	15.09
6	8.56	10.64	12.59	14.45	16.81
7	9.80	12.02	14.07	16.01	18.48
8	11.03	13.36	15.51	17.54	20.09
9	12.24	14.68	16.92	19.02	21.67
10	13.44	15.99	18.31	20.48	23.21

Using the table, we find that the critical value for a 0.05 significance level with $df = 2$ is 5.99. That means that 95 times out of 100, a survey that agrees with a sample will have a χ^2 value of 5.99 or less. If our chi-square value is *greater* than 5.995, then the measurements we took only occur 5 or fewer times out of 100, or the null hypothesis is incorrect. Our chi-square statistic is only **0.7533**, so we will ***not reject*** the null hypothesis.

- **Step 4:** Interpret the results.

Since our chi-square statistic was less than the critical value, we do not reject the null hypothesis, and we can say that our survey data does support the data from the APPA.

Example B

Rachel told Eric that the reason her car insurance is less expensive is that female drivers get in fewer accidents than male drivers. Specifically, she says that male drivers are held responsible in 65% of accidents involving drivers under 23. If Eric does some research of his own and discovers that 46 out of the 85 accidents he investigates involve male drivers, does his data support or refute Rachel's hypothesis?

Solution

- **Step 1:** Clearly state the null and alternative hypotheses.

  H_0: *Male drivers are responsible for 65% of accidents and female drivers are responsible for 35%.*

  H_1: *The data do not match the proposed model.*

- **Step 2:** Identify an appropriate test and significance level.

  Again, a Chi-Square goodness of fit test is appropriate. In the absence of a stated significance level in the problem, we assume the default 0.05.

- **Step 3:** Analyze sample data.

  Create a table to organize data and compare the observed data to the expected data:

TABLE 14.4:

	Male Drivers	Female Drivers	TOTAL
Observed	46	39	85
Expected			

To identify the expected values, multiply the expected % by the total number observed:

TABLE 14.5:

	Male Drivers	Female Drivers	TOTAL
Observed	46	39	85
Expected	$0.65 \times 85 = 55.25$	$0.35 \times 85 = 29.75$	85

To calculate our chi-square statistic, we need to sum the squared differences between each observed and expected value divided by the expected value:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

$$\chi^2 = \frac{(46 - 55.25)^2}{55.25} + \frac{(39 - 29.75)^2}{29.75}$$

$$\chi^2 = \frac{(-9.25)^2}{55.25} + \frac{(9.25)^2}{29.75}$$

$$\chi^2 = \frac{85.5625}{55.25} + \frac{85.5625}{29.75}$$

$$\chi^2 = 1.5486 + 2.8760$$

$$\chi^2 = 4.4246$$

Now that we have our chi-square statistic, we need to compare it to the chi-square critical value for 0.05 with *one degree of freedom*, since we have two categories. Using the table, we find the critical value to be 3.84. The critical value indicates that only 0.05, or 5%, of values would be as high as 3.84. If the χ^2 of our data is greater than 3.84, then fewer than 5 times out of 100 would we expect to get that result if the null hypothesis is true.

- **Step 4:** Interpret your results.

  Our calculated data value of $\chi^2 = 4.4246$ is greater than the 0.05 significance level critical value of 3.84, so we reject the null hypothesis. The data that Eric observed does not support the distribution that Rachel claimed.

Example C

The online car magazine "*Camaro5.com*" claims that, given a choice between a Ford Mustang or Chevy Camaro, 51% of readers will choose a Camaro. Ellen is a Mustang lover and decides to do some research. If Ellen collects the data below, does her data support the magazine's claim?

Data: Mustang owners: 28, Camaro owners: 34

Solution

- **Step 1:** Clearly state the null and alternative hypotheses.

  H_0: 51% of readers prefer a Camero to a Mustang.

  H_1: Readers do not have the proposed preference.

- **Step 2:** Identify an appropriate test and significance level.

  A Chi-Square Goodness of Fit test is appropriate. In the absence of a stated significance level in the problem, we assume the default 0.05.

- **Step 3:** Analyze sample data.

  We will start by creating a table to organize our data:

TABLE 14.6:

	Mustang	Camaro	TOTAL
Observed	28	34	62
Expected	$0.49 \times 62 = 30.4$	$0.51 \times 62 = 31.6$	62

Now we can calculate our chi statistic:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

$$\chi^2 = \frac{(28 - 30.4)^2}{30.4} + \frac{(34 - 31.6)^2}{31.6}$$

$$\chi^2 = \frac{(-2.4)^2}{30.4} + \frac{(2.4)^2}{31.6}$$

$$\chi^2 = .3718$$

The chi-square critical value for $df = 1$ and a significance level of 0.05 is 3.8414 (the same as in Example B).

- **Step 4:** Interpret your results.

Our calculated data value of $\chi^2 = 0.3718$ is significantly less than the 0.05 significance level critical value of 3.8141, so we fail to reject the null hypothesis. This means that, unfortunately for Ellen, her research did not allow her to deny the claim that Camaros are more popular.

Features of the Goodness-of-Fit Test

As mentioned, the goodness-of-fit test is used to determine patterns of distinct categorical variables. The test requires that the data are obtained through a random sample. The number of *degrees of freedom* associated with a particular chi-square test is equal to the number of categories minus one. That is, $df = c - 1$.

Example D

So for a test which involves four categories, we calculate the degrees of freedom as follows:

$$df = \text{number of categories} - 1$$
$$3 = 4 - 1$$

There are many situations that use the goodness-of-fit test, including surveys, taste tests, and analysis of behaviors. Interestingly, goodness-of-fit tests are also used in casinos to determine if there is cheating in games of chance, such as cards or dice. For example, if a certain card or number on a die shows up more than expected (a high observed frequency compared to the expected frequency), officials use the goodness-of-fit test to determine the likelihood that the player may be cheating or that the game may not be fair.

Lesson Summary

We use the chi-square test to examine patterns between categorical variables, such as genders, political candidates, locations, or preferences.

To test for significance, it helps to make a table detailing the observed and expected frequencies of the data sample. Using the standard chi-square distribution table, we are able to create criteria for accepting the null or alternative hypotheses for our research questions.

To test the null hypothesis, it is necessary to calculate the chi-square statistic, χ^2. To calculate the chi-square statistic, we use the following formula:

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

where:

χ^2 is the chi-square test statistic.

O is the observed frequency value for each event.

E is the expected frequency value for each event.

Using the chi-square statistic and the level of significance, we are able to determine whether to reject or fail to reject the null hypothesis and write a summary statement based on these results.

Vocabulary

A **chi-square** statistic is a derived value used in a **chi-square test** to calculate the probability that a given distribution is a good fit for observed data.

The **degrees of freedom** of a variable are the number of values in the final calculation of a statistic that are free to vary. The degrees of freedom are calculated as $n-1$, where n is the number of samples or categories in the variable.

Guided Practice

Questions 1-5 refer to the following data:

Tuscany claims that 70% of local pet owners own a dog, and 30% own a cat. Sayber decides to test her claim and learns that 23 of the 40 people he asks own dogs, and 17 own cats.

1. What kind of test could you use to see if Sayber's data supports Tuscany's claim?
2. What would be the null and alternative hypotheses?
3. What would be the expected values of dog and cat owners?
4. What is the chi-square statistic of the observed data?
5. Assuming a 0.1 significance level, does Sayber's data support Tuscany's claim?

Solutions

1. A chi-square goodness of fit test would be appropriate because this is a claim about a distribution of a single categorical variable.
2. The null hypothesis, H_0, would be that 70% of pet owners own a dog and 30% own a cat; the alternative hypothesis would be that this model does not fit the data.
3. The expected number of dog owners, according to Tuscany's claim, would be 70% of the 40 people that Sayber polled, or 28 dog owners. The expected number of cat owners would be 30% of the 40 people polled, or 12.
4. The χ^2 statistic is the sum of the squared differences between the observed and expected values, divided by the expected values:

$$\chi^2 = \frac{(23-28)^2}{28} + \frac{(17-12)^2}{12}$$
$$= 0.893 + 2.083$$
$$\chi^2 = 2.976$$

5. Df = 2-1 = 1. The critical value of chi-squared for 1 degree of freedom at a significance level of 0.1 is 2.71. Since the chi-square statistic we calculated is 2.976, and is therefore more extreme than the critical value, we may *reject the hypothesis*, and say that Sayber's data does not support Tuscany's claim.

More Practice

Questions 1-5 refer to the following:

Evan claims that 15% of computer gamers have played "Team Fortress 2", and 35% have played "World of Warcraft". Evan's brother is skeptical of those figures and decides to do some research. He discovers that 60 of the 200 computer gamers he polls have played "Team Fortress 2", and 90 have played "World of Warcraft".

1. Create a table to organize the data and prepare for hypothesis testing.
2. What sort of test would be appropriate to determine if the observed data supports Evan's claim?
3. What would be H_0 *and* H_1?
4. What would be the χ^2 statistic for the observed data?
5. How many degrees of freedom are there in the variable "played game"?
6. Assuming a significance level of 0.05, what is the χ^2 critical value?
7. Does the observed data support Evan's claim? Explain your findings.

Questions 8-15 refer to the following:

Mack claims that 84% of street racers drive import cars, and 16% drive domestic muscle cars. Abbi likes domestic cars and thinks Mack is overstating the percentage of imports, so she does some research of her own and finds that 57 of the street racers she interviewed drive imports, and 31 drive American muscle.

8. Create a table to organize the data and prepare for hypothesis testing.
9. What sort of test would be appropriate to determine if the observed data supports Mack's claim?
10. What would be H_0 *and* H_1?
11. What would be the χ^2 statistic for the observed data?
12. How many degrees of freedom are there in the variable "played game"?
13. Assuming a significance level of 0.10, what is the χ^2 critical value?
14. Does the data indicate that Abbi should reject, or fail to reject H_0?
15. Interpret your results.

14.2 Test of Independence

Learning Objectives

- Understand how to draw data needed to perform calculations when running the chi-square test from contingency tables.
- Run the test of independence to determine whether two variables are independent or not.

Introduction

The chi-square test can be used to both estimate how closely an observed distribution matches an expected distribution (the goodness-of-fit test) and to estimate whether two random variables are independent of one another (the test-of-independence). In this lesson, we will examine the test of independence in greater detail.

The chi-square test of independence is used to assess if two factors are related. This test is often used in social science research to determine if factors are independent of each other. For example, we would use this test to determine relationships between voting patterns and race, income and gender, and behavior and education.

In general, when running the test of independence, we ask, "Is Variable X independent of Variable Y?" It is important to note that this test does not test how the variables are related, just simply whether or not they are independent of one another. For example, while the test of independence can help us determine if income and gender are independent, it cannot help us assess how one category might affect the other.

Test of Independence

A chi-square (χ^2) test can be used to determine if observed data indicates that two variables are dependent in much the same way that the test can be used to determine goodness of fit.

Just as with a goodness of fit test, we will calculate expected values, calculate a chi-square statistic, and compare it to the appropriate chi-square value from a reference to see if we should reject $H0$, which is that the *variables are not related*. Formally the hypothesis statements for the chi-square Test-of-Independence are:

H_0 : There is no association between the two categorical variables

H_A : There is an association (the two variables are not independent)

In fact, the only major difference in process between a goodness of fit test and a test of independence is how we calculate the expected values, as you will see in Example A.

Just for reference:

- The formula to calculate chi-square is:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

- The formula for calculating expected values in a test of independence is:

$$expected\ cell\ value = \frac{C \times R}{n}$$

Where C is the observed column total for the cell, R is the observed row total for the cell, and n is the total number of samples. (see Example A for clarification of the use of the formula)

- The degrees of freedom in a test of independence are calculated as:

$$df = (rows - 1)(columns - 1)$$

Contingency tables can help us frame our hypotheses and solve problems. Often, we use contingency tables to list the variables and observational patterns that will help us to run a chi-square test. For example, we could use a contingency table to record the answers to phone surveys or observed behavioral patterns.

Finally, let's take a look at the last piece of the test, the assumptions of the test:

Assumptions of the chi-square test:

- Random sample.
- Independent observations for the sample (one observation per subject).
- All expected counts greater than one.
- No more than 20% of cells with an expected count less than five.

Example A

Let's look at a contingency table with some count data to help us understand how a test of independence works. Here, a total of 336 individuals identified both their happiness level and their belief in an afterlife.

TABLE 14.7:

Belief in after-life		Happiness level:			
	Not Happy	**Pretty Happy**	**Very Happy**	**Total**	
Yes	30	145	95	270	
No	10	44	12	66	
Total	40	189	107	336	

Let's calculate the expected values in each cell of the table:

[yes + not happy] $= \frac{270*40}{336} = 32.14$

[yes + pretty happy] $= \frac{270*189}{336} = 151.88$

[yes + very happy] $= \frac{270*107}{336} = 85.98$

[no + not happy] $= \frac{66*40}{336} = 7.86$

[no + pretty happy] $= \frac{66*189}{336} = 37.12$

[no + very happy] $= \frac{66*107}{336} = 21.02$

We'll now use the data from the above contingency table, and the expected values, to statistically test if happiness is associated with a belief in the afterlife.

- **Hypothesis Step 1:** Clearly state the Null and Alternative Hypothesis.

   H_0 : There is no association between happiness and the belief in the afterlife

   H_A : There is an association (the two variables are not independent)

- **Hypothesis Step 2**: Identify the appropriate significance level and confirm the test assumptions.

   We'll use the standard significance level of 0.05. We were told the survey was random, and we'll assume that a single subject is not living in two places at once. From our expected counts that we calculated, we can confirm that the last two assumptions have been met.

- **Hypothesis Step 3 :** Analyze the data.

   Using the chi-square statistic:

   $$\chi^2 = \Sigma \frac{(\text{observed - expected})^2}{\text{expected}}$$

   And using the chi-square formula:

   $$\chi^2 = \frac{(30-32.14)^2}{32.14} + \frac{(145-151.88)^2}{151.88} + \frac{(95-85.98)^2}{85.98} + \frac{(10-7.86)^2}{7.86} + \frac{(44-37.12)^2}{37.12} + \frac{(12-21.02)^2}{21.02} = 7.13$$

   We already know we have 2 degrees of freedom, and looking up 2 df on the chi-square distribution table we find a critical value of 5.99.

- **Hypothesis Step 4:** Interpret your results.

   We find that our calculated chi-square value is greater than our critical value, so we can reject our Null hypothesis. We conclude that happiness is not independent of a belief in the afterlife.

Example B

Rachel claims that girls take more black and white and color photographs than boys, but Jack (who is a photographer) is skeptical. If Jack collects the following data, would it be correct to say that he should reject Rachel's claim that gender affects tendency to take photographs?

TABLE 14.8:

	Black/White	Color	TOTAL
Female	72	489	**561**
Male	48	530	**578**
TOTAL	**120**	**1019**	**1139**

Solution

The question here is whether gender affects tendency to take more photographs, or, in other words, are gender and photograph-taking tendency dependent?

To run a chi-squared test, we need to know the expected value for each of the four cells containing observations. In a test for independence, this is calculated with the formula: *expected cell value* $= \frac{C \times R}{n}$.

1. The upper-left cell, female X black/white:

$$
\begin{aligned}
expected\ cell\ value &= \frac{C \times R}{n} \\
&= \frac{(column\ total) \times (row\ total)}{total\ number\ of\ observations} \\
&= \frac{120 \times 561}{1139} \\
expected\ cell\ value &= 59.1
\end{aligned}
$$

2. The cell below that, male X black/white:

$$
\begin{aligned}
expected\ cell\ value &= \frac{C \times R}{n} \\
&= \frac{(column\ total) \times (row\ total)}{total\ number\ of\ observations} \\
&= \frac{120 \times 578}{1139} \\
expected\ cell\ value &= 60.9
\end{aligned}
$$

3. Top-right cell, female X color:

$$
\begin{aligned}
expected\ cell\ value &= \frac{C \times R}{n} \\
&= \frac{(column\ total) \times (row\ total)}{total\ number\ of\ observations} \\
&= \frac{1019 \times 561}{1139} \\
expected\ cell\ value &= 501.9
\end{aligned}
$$

4. Bottom-right cell, male X color:

$$expected\ cell\ value = \frac{C \times R}{n}$$
$$= \frac{(column\ total) \times (row\ total)}{total\ number\ of\ observations}$$
$$= \frac{1019 \times 578}{1139}$$
$$expected\ cell\ value = 517.1$$

Now we can add the expected values to our initial table, placing the expected value for each cell in parentheses:

TABLE 14.9:

	Black/White	Color	TOTAL
Female	72 (59.1)	489 (501.9)	561
Male	48 (60.9)	530 (517.1)	578
TOTAL	120	1019	1139

Now we can calculate our χ^2 statistic as before, using: $\chi^2 = \sum \frac{(observed - expected)^2}{expected}$ and each of the four values in the body of the table:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$
$$\chi^2 = \frac{(72 - 59.1)^2}{59.10} + \frac{(48 - 60.1)^2}{60.89} + \frac{(489 - 501.9)^2}{501.89} + \frac{(530 - 517.1)^2}{517.10}$$
$$\chi^2 = \frac{(12.9)^2}{59.10} + \frac{(-12.9)^2}{60.89} + \frac{(-12.9)^2}{501.89} + \frac{(12.9)^2}{517.10}$$
$$\chi^2 = \frac{166.41}{59.10} + \frac{166.41}{60.89} + \frac{166.41}{501.89} + \frac{166.41}{517.10}$$
$$\chi^2 = 2.82 + 2.73 + .33 + .32$$
$$\chi^2 = 6.20$$

To see if our χ^2 statistic is greater or less than the critical value at the default significance level of 0.05, we need the number of *degrees of freedom*: $df = (rows - 1)(columns - 1)$

$$df = (2 - 1)(2 - 1)$$
$$df = 1$$

Using our chi-squared critical value reference, we find that the critical value for 0.05 with $df = 1$ is 3.8414.

Finally, we compare our calculated chi-squared value of 6.2 to the critical value of 3.8414 and determine that since $6.2 > 3.8414$, we can *reject H_0*, in other words, we reject the independence of the variables . **The observed data indicates that there is a gender bias on picture-taking tendency.**

Vocabulary

A **chi-squared** (χ^2) statistic is calculated using $\chi^2 = \sum \frac{(observed - expected)^2}{expected}$, and may be used to evaluate variable independence.

The **expected value**, as used in a chi-square test, is the value you would expect to get if the null hypothesis is correct.

Guided Practice

Questions 1-5 refer to the following:

Kato claims that single people prefer different pizzas than married people do. Kato's brother doesn't think that is true, so he conducts some research of his own, and collects the data below.

TABLE 14.10:

	Pepperoni	**Sausage**	**Cheese**	**TOTAL**
Single	29	12	61	102
Married	8	47	56	111
TOTAL	37	59	117	213

1. Fill in the expected values of the 6 cells in the body of the table using parenthesis.
2. What is the value of χ^2?
3. How many degrees of freedom are there?
4. If we plan to test the claim, what are H_0 *and* H_1?
5. Assuming a significance level of 0.05, does the observed data support Kato's claim?

Solutions

1. Completed table, using *expected cell value* $= \frac{C \times R}{n}$:

TABLE 14.11:

	Pepperoni	**Sausage**	**Cheese**	**TOTAL**
Single	29 (17.71)	12 (28.25)	61 (56.02)	102
Married	8 (19.28)	47 (30.74)	56 (60.97)	111
TOTAL	37	59	117	213

For example, Expected (single, pepperoni) = 102(37)/213 = 17.718

2. Using $\chi^2 = \sum \frac{(observed - expected)^2}{expected}$:

$$\chi^2 = \frac{(29 - 17.718)^2}{17.718} + \frac{(12 - 28.254)^2}{28.254} + \frac{(61 - 56.028)^2}{56.028} + \frac{(8 - 19.282)^2}{19.282} + \frac{(47 - 30.746)^2}{30.746} + \frac{(56 - 60.972)^2}{60.972}$$

$$\chi^2 = 7.184 + 9.351 + 0.441 + 6.601 + 8.593 + 0.405$$

$$\chi^2 = 32.575$$

3. $df = (3 - 1)(2 - 1) = 2$
4. H_0: Marital status and pizza type are not associated, H_1: Marital type and pizza type are associated.
5. The critical value for χ^2 with $2df$ at 0.05 significance is 5.99. Since our calculated value of $\chi^2 = 32.575$, and $32.575 > 5.99$ we can **reject the null hypothesis.** There is enough evidence to suggest that marital status is associated with pizza preferences, therby supporting Kato's claim.

More Practice

1. What use of the χ^2 statistic is used in this lesson?
2. How is an expected value calculated, in the context of this lesson?
3. How are degrees of freedom calculated in a chi-square test of independence?
4. What type of table is commonly used to organize information for a chi-square test of independence?
5. What is the default level of significance for a chi-squared test of independence?

Questions 6–10 refer to the following breakdown of favorite flavor by gender:

TABLE 14.12:

	Cherry	Lemon	Strawberry	Other	TOTAL
Male	13	11	7	13	**44**
Female	15	18	11	5	**49**
TOTAL	**28**	**29**	**18**	**18**	**93**

6. Fill in the expected values of the 6 cells in the body of the table using parenthesis.
7. What is the value of χ^2?
8. How many degrees of freedom are there?
9. If we plan to test the claim that gender affects favorite flavor, what are H_0 and H_1?
10. Assuming a significance level of 0.05, does the observed data indicate that we reject or fail to reject H_0?

Questions 10–14 refer to the following:

Are hamburger cooking preferences dependent on gender? 1087 people were asked their preference among three ways of cooking burgers, grilling, frying, and broiling. The men stated their preferences as: 137: grilling, 193: frying, and 212: broiling. The women were distributed as: 110: grilling, 215: frying, and 220: broiling.

11. Create a contingency table to organize the information.
12. What is the value of χ^2?
13. How many degrees of freedom are there?
14. If we plan to test the claim that gender affects cooking preference, what are H_0 and H_1?