The Chi-Square Distribution: Teacher's Guide*

Susan Dean
Barbara Illowsky, Ph.D.

Abstract

This module is the complementary teacher's guide for the "The Chi-Square Distribution" chapter of the Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

This chapter is concerned with three chi-square applications: goodness-of-fit; independence; and single variance. We rely on technology to do the calculations, especially for goodness-of-fit and for independence. However, the first example in the chapter (the number of absences in the days of the week) has the student calculate the chi-square statistic in steps. The same could be done for the chi-square statistic in a test of independence.

The chi-square distribution generally is skewed to the right. There is a different chi-square curve for each df. When the df's are 90 or more, the chi-square distribution is a very good approximation to the normal. For the chi-square distribution, μ = the number of df's and σ = the square root of twice the number of df's.

Goodness-of-Fit Test

A goodness-of-fit hypothesis test is used to determine whether or not data "fit" a particular distribution.

Example 1

In a past issue of the magazine *GEICO Direct*, there was an article concerning the percentage of teenage motor vehicle deaths and time of day. The following percentages were given from a sample.

*Version 1.11: Apr 5, 2010 7:46 pm -0500

†http://creativecommons.org/licenses/by/2.0/

Time of Day Percentage of Motor Vehicle Deaths

Time of Day	Death Rate
12 a.m. to 3 a.m.	17%
3 a.m. to 6 a.m.	8%
6 a.m. to 9 a.m.	8%
9 a.m. to 12 noon	6%
12 noon to 3 p.m.	10%
3 p.m. to 6 p.m.	16%
6 p.m. to 9 p.m.	15%
9 p.m. to 12 a.m.	19%

Table 1

For the purpose of this example, suppose another sample of 100 produced the same percentages. We hypothesize that the data from this new sample fits a uniform distribution. The level of significance is 1% ($\alpha = 0.01$).

- H_o: The number of teenage motor vehicle deaths fits a uniform distribution.
- H_a : The number of teenage motor vehicle deaths does not fit a uniform distribution.

The distribution for the hypothesis test is X_7^2

The table contains the observed percentages. For the sample of 100, the observed (O) numbers are 17, 8, 8, 6, 10, 16, 15 and 19. The expected (E) numbers are each 12.5 for a uniform distribution (100 divided by 8 cells). The chi-square test statistic is calculated using

$\sum_8 \frac{(0-E)^2}{E}$
$= \frac{(17-12.5)^2}{12.5} + \frac{(8-12.5)^2}{12.5} + \frac{(8-12.5)^2}{12.5} + \frac{(6-12.5)^2}{12.5} + \frac{(10-12.5)^2}{12.5} + \frac{(16-12.5)^2}{12.5} + \frac{(15-12.5)^2}{12.5} + \frac{(19-12.5)^2}{12.5}$
$= 13.6$

If you are using the TI-84 series graphing calculators, ON SOME OF THEM there is a function in STAT TESTS called x^2 GOF-Test that does the goodness-of-fit test. You first have to enter the observed numbers in one list (enter as whole numbers) and the expected numbers (uniform implies they are each 12.5) in a second list (enter 12.5 for each entry: 100 divided by 8 = 12.5). Then do the test by going to x^2 GOF-Test.

If you are using the TI-83 series, enter the observed numbers in list1 and the expected numbers in list2 and in list3 (go to the list name), enter (list1-list2)^2/list2. Press enter. Add the values in list3 (this is the test statistic). Then go to 2nd DISTR x^2cdf. Enter the test statistic (13.6) and the upper value of the area (10^99) and the degrees of freedom (7).

Probability Statement: $P\left(x^2 > 13.6\right) = 0.0588$

(Always a right-tailed test)

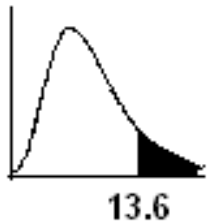

Figure 1: p-value $= 0.0588$

Since $\alpha < p$-value $(0.01 < 0.0588)$, we do not reject H_o.

We conclude that there is not sufficient evidence to reject the null hypothesis. It appears that the number of teenage motor vehicle deaths fits a uniform distribution. It does not matter what time of the day or night it is. Teenagers die from motor vehicle accidents equally at any time of the day or night. However, if the level of significance were 10%, we would reject the null hypothesis and conclude that the distribution of deaths does not fit a uniform distribution.

A **test of independence** compares two factors to determine if they are independent (i.e. one factor does not affect the happening of a second factor).

Example 2

The following table shows a random sample of 100 hikers and the area of hiking preferred.

Hiking Preference Area

Gender	The Coastline	Near Lakes and Streams	On Mountain Peaks
Female	18	16	11
Male	16	25	14

Table 2: The two factors are gender and preferred hiking area.

- H_o: Gender and preferred hiking area are independent.
- H_a: Gender and preferred hiking area are not independent

The distribution for the hypothesis test is x_2^2.
The df's are equal to: $(\text{rows} - 1)(\text{columns} - 1) = (2 - 1)(3 - 1) = 2$
The chi-square statistic is calculated using $\sum_{(2-3)} \frac{(0-E)^2}{E}$
Each expected (E) value is calculated using $\frac{(\text{rowtotal})(\text{columntotal})}{\text{totalsurveyed}}$
The first expected value (female, the coastline) is $\frac{45 \cdot 34}{100} = 15.3$
The expected values are: 15.3, 18.45, 11.25, 18.7, 22.55, 13.75
The chi-square statistic is:
$\sum_{(2-3)} \frac{(0-E)^2}{E} =$
$\frac{(18-15.3)^2}{15.3} + \frac{(16-18.45)^2}{18.45} + \frac{(11-11.15)^2}{11.25} + \frac{(16-18.7)^2}{18.7} + \frac{(25-22.55)^2}{22.55} + \frac{(14-13.75)^2}{13.75}$
$= 1.47$

Calculator Instructions

The TI-83/84 series have the function x^2-Test in STAT TESTS to preform this test. First, you have to enter the observed values in the table into a matrix by using 2nd MATRIX and EDIT [A]. Enter the values and go to x^2-Test. Matrix [B] is calculated automatically when you run the test.

Probability Statement: p-value $= 0.4800$ (A right-tailed test)

Figure 2: p-value $= 0.4800$

Since α is less than 0.05, we do not reject the null.

There is not sufficient evidence to conclude that gender and hiking preference are not independent.

Sometimes you might be interested in how something varies. A **test of a single variance** is the type of hypothesis test you could run in order to determine variability.

Example 3

A vending machine company which produces coffee vending machines claims that its machine pours an 8 ounce cup of coffee, on the average, with a standard deviation of 0.3 ounces. A college that uses the vending machines claims that the standard deviation is more than 0.3 ounces causing the coffee to spill out of a cup. The college sampled 30 cups of coffee and found that the standard deviation was 1 ounce. At the 1% level of significance, test the claim made by the vending machine company.

Solution

$H_o : \sigma^2 = (0.3)^2$ $H_a : \sigma^2 > (0.3)^2$

The distribution for the hypothesis test is x^2_{29} where df $= 30 - 1 = 29$.

The test statistic $x^2 = \frac{(n-1)\cdot s^2}{\sigma^2} = \frac{(30-1)\cdot 1^2}{0.3^2} = 322.22$

Probability Statement: $P\left(x^2 > 322.22\right) = 0$

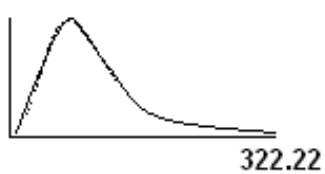

Figure 3: p-value $= 0$

Since $a > p$-value $(0.01 > 0)$, reject H_o.

There is sufficient evidence to conclude that the standard deviation is more than 0.3 ounces of coffee. The vending machine company needs to adjust their machines to prevent spillage.

Assign Practice

Have the students do the Practice 1[1], Practice 2[2], and Practice 3[3] in class collaboratively.

Assign Homework

Assign Homework [4]. Suggested homework: 3, 5, 7 (GOF), 9, 13, 15 (Test of Indep.), 17, 19, 23 (Variance), 24 - 37 (General)

[1] "The Chi-Square Distribution: Practice 1" <http://cnx.org/content/m17054/latest/>
[2] "The Chi-Square Distribution: Practice 2" <http://cnx.org/content/m17056/latest/>
[3] "The Chi-Square Distribution: Practice 3" <http://cnx.org/content/m17053/latest/>
[4] "The Chi-Square Distribution: Homework" <http://cnx.org/content/m17028/latest/>