The Chi-Square Distribution: Test of Independence*

Susan Dean
Barbara Illowsky, Ph.D.

Abstract

This module describes how the chi-square distribution can be used to test for independence.

Tests of independence involve using a **contingency table** of observed (data) values. You first saw a contingency table when you studied probability in the Probability Topics[1] chapter.

The test statistic for a test of independence is similar to that of a goodness-of-fit test:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E} \tag{1}$$

where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Chapter 3. As a review, consider the following example.

NOTE: The expected value for each cell needs to be at least 5 in order to use this test.

Example 1

Suppose A = a speeding violation in the last year and B = a cell phone user while driving. If A and B are independent then $P(A\ AND\ B) = P(A)\,P(B)$. $A\ AND\ B$ is the event that a driver received a speeding violation last year and is also a cell phone user while driving. Suppose, in a study of drivers who received speeding violations in the last year and who uses cell phones while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 were cell phone users while driving and 450 were not.

*Version 1.12: Jun 19, 2012 8:12 pm +0000
†http://creativecommons.org/licenses/by/3.0/
[1]"Probability Topics: Introduction" <http://cnx.org/content/m16838/latest/>

Let y = expected number of drivers that use a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A\ AND\ B) = P(A)\,P(B)$. By substitution,

$\frac{y}{755} = \frac{70}{755} \cdot \frac{305}{755}$

Solve for y : $y = \frac{70 \cdot 305}{755} = 28.3$

About 28 people from the sample are expected to be cell phone users while driving and to receive speeding violations.

In a test of independence, we state the null and alternate hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternate hypothesis states that they are **not independent (dependent)**. If we do a test of independence using the example above, then the null hypothesis is:

H_o: Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to be cell phone users while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, like goodness-of-fit.

The degrees of freedom for the test of independence are:

df = (number of columns - 1)(number of rows - 1)

The following formula calculates the **expected number** (E):

$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}}$

Example 2

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. The following table is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Number of Hours Worked Per Week by Volunteer Type (Observed)

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Table 1: The table contains **observed (O)** values (data).

Problem

Are the number of hours volunteered **independent** of the type of volunteer?

Solution

The **observed table** and the question at the end of the problem, "Are the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are **number of hours volunteered** and **type of volunteer**. This test is always right-tailed.

H_o: The number of hours volunteered is **independent** of the type of volunteer.

H_a: The number of hours volunteered is **dependent** on the type of volunteer.

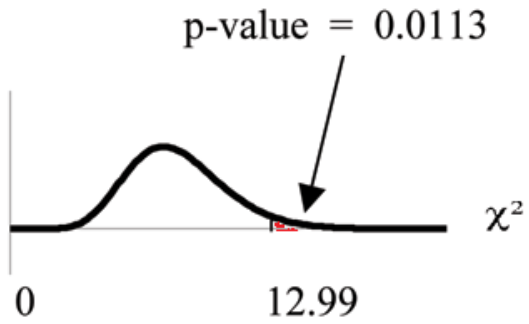
The expected table is:

Number of Hours Worked Per Week by Volunteer Type (Expected)

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

Table 2: The table contains **expected** (E) values (data).

For example, the calculation for the expected frequency for the top left cell is
$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}} = \frac{255 \cdot 298}{839} = 90.57$
Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)
Distribution for the test: χ^2_4
df = (3 columns − 1) (3 rows − 1) = (2) (2) = 4
Graph:



Probability statement: p-value = $P\left(\chi^2 > 12.99\right) = 0.0113$
Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. p-value = 0.0113. $\alpha >$ p-value.
Make a decision: Since $\alpha >$ p-value, reject H_o. This means that the factors are not independent.
Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the above example, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

NOTE: Calculator instructions follow.

TI-83+ and TI-84 calculator: Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row from Example 11-6. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:`χ`2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 12.9909 and the p-value = 0.0113. Do the procedure a second time but arrow down to `Draw` instead of `calculate`.

Example 3

De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. The table shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School vs. Anxiety Level

Need to Succeed in School	High Anxiety	Med-high Anxiety	Medium Anxiety	Med-low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

Table 3

Problem 1

How many high anxiety level students are expected to have a high need to succeed in school?

Solution

The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

Problem 2

If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

Solution

The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

Problem 3

a. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} =$

b. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about:

Glossary

Definition 1: Contingency Table
The method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other. The table provides an easy way to calculate conditional probabilities.