11.5: Test of Independence

Tests of independence involve using a **contingency table** of observed (data) values. The test statistic for a **test of independence** is similar to that of a goodness-of-fit test:

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

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 Table 3.1 earlier. As a review, consider the following example.

Note

The expected value inside each cell needs to be at least five in order for you to use this test.

Example 11.5.1

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

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

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

$$\frac{y}{755} = \left(\frac{70}{755}\right)\left(\frac{305}{755}\right)$$

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

About 28 people from the sample are expected to use cell phones while driving and to receive speeding violations.

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

H_0: Being a cell phone user while driving and receiving a speeding violation are independent events; in other words, they have no effect on each other.

If the null hypothesis were true, we would expect about 28 people to use cell phones 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, as it is in a goodness-of-fit.

The number of degrees of freedom for the test of independence is:

$df = (\text{number of columns} - 1)(\text{number of rows} - 1)$

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

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

Try It 11.5.1

A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. Ninety-seven of the 300 surveyed were on the honor roll, while 203 were not. If we assume being a music student and being on the honor roll are independent events, what is the expected number of music students who are also on the honor roll?

Exercise 11.5.1

A volunteer group, provides from one to nine hours each week with disabled senior citizens. The program recruits among community college students, four-year college students, and nonstudents. In Table 11.14 is a **sample** of the adult volunteers and the number of hours they volunteer per week.

The table contains **observed (O)** values (data).

Table 11.14 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

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

Answer

Solution 11.9

The **observed table** and the question at the end of the problem, "Is 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_0: 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 result are in Table 11.15.

The table contains **expected** (E) values (data).

Table 11.15 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

For example, the calculation for the expected frequency for the top left cell is

$$E = \frac{(\text{ row total })(\text{ column total })}{\text{total number surveyed}} = \frac{(255)(298)}{839} = 90.57$$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

$df = (3 \text{ columns } -1)(3 \text{ rows } -1) = (2)(2) = 4$

Graph:

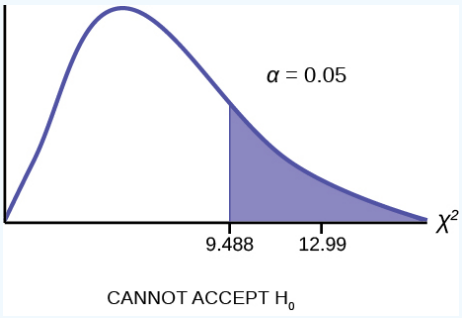

Figure 11.8

The graph of the Chi-square shows the distribution and marks the critical value with four degrees of freedom at 95% level of confidence, $\alpha = 0.05$, 9.488. The graph also marks the calculated χ^2_c test statistic of 12.99. Comparing the test statistic with the critical value, as we have done with all other hypothesis tests, we reach the conclusion.

Make a decision: Because the calculated test statistic is in the tail we cannot accept H_0. 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 example in Table 11.15, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

Try It 11.5.2

The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. Table 11.16 shows the results:

Table 11.16

Industry sector	2000	2010	2020	Total
Nonagriculture wage and salary	13,243	13,044	15,018	41,305
Goods-producing, excluding agriculture	2,457	1,771	1,950	6,178
Services-providing	10,786	11,273	13,068	35,127
Agriculture, forestry, fishing, and hunting	240	214	201	655
Nonagriculture self-employed and unpaid family worker	931	894	972	2,797
Secondary wage and salary jobs in agriculture and private household industries	14	11	11	36

Industry sector	2000	2010	2020	Total
Secondary jobs as a self-employed or unpaid family worker	196	144	152	492
Total	27,867	27,351	31,372	86,590

We want to know if the change in the number of jobs is independent of the change in years. State the null and alternative hypotheses and the degrees of freedom.

Exercise 11.5.2

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. Table 11.17 shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Table 11.17 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

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

Answer

Solution 11.10

a. 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.

b. 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?

Answer

Solution 11.10

b. 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.

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

Answer

Solution 11.10

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = 8.19$

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

Answer

Solution 11.10

d. 8

This page titled 11.5: Test of Independence is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **11.4: Test of Independence** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.