

## 10.3: Test for Independence

Use the chi-square test for independence to test the independence of two categorical variables. Remember, qualitative data is collected on individuals that are categories or names. Then you would count how many of the individuals had particular qualities. An example is that there is a theory that there is a relationship between breastfeeding and having autism spectrum disorder (ASD). To determine if there is a relationship, researchers could collect the time-period that a mother breastfed her child and if that child was diagnosed with ASD. Then you would have a table containing this information. Now you want to know if each cell is independent of each other cell. Remember, independence says that one event does not affect another event. Here it means that having ASD is independent of being breastfed. What you really want is to see if they are dependent (not independent). In other words, does one affect the other? If you were to do a hypothesis test, this is your alternative hypothesis and the null hypothesis is that they are independent. There is a hypothesis test for this and it is called the **chi-square test for independence**.

There is only a right-tailed test for testing the independence between two variables:

$H_0$  : Variable 1 and Variable 2 are independent (unrelated).

$H_1$  : Variable 1 and Variable 2 are dependent (related).

Finding the test statistic involves several steps. First, the data is collected, counted, and then organized into a contingency table. These values are known as the observed frequencies, and the symbol for an observed frequency is  $O$ . Total each row and column.

The null hypothesis is that the two variables are independent. If two events are independent then  $P(B) = P(B|A)$  and we can use the multiplication rule for independent events, to calculate the probability that variable  $A$  and  $B$  as the  $P(A \text{ and } B) = P(A) \cdot P(B)$ . Remember in a hypothesis test, you assume that  $H_0$  is true, the two variables are assumed to be independent.

$$\begin{aligned}
 P(A \text{ and } B) &= P(A) \cdot P(B) \text{ if } A \text{ and } B \text{ are independent} \\
 &= \frac{\text{Number of ways A can happen}}{\text{Total number of individuals}} \cdot \frac{\text{Number of ways B can happen}}{\text{Total number of individuals}} \\
 &= \frac{\text{Row Total}}{n} \cdot \frac{\text{Column Total}}{n}
 \end{aligned}$$

Variable 2	Variable 1		Total
	B	Not B	
A	P(A and B)	P(A and Not B)	Row Total for A
Not A	P(Not A and B)	P(Not A and Not B)	Row Total for Not A
Total	Column Total for B	Column Total for Not B	$n$

Now you want to find out how many individuals you expect to be in a certain cell. To find the expected frequencies, you just need to multiply the probability of that cell times the total number of individuals. Do not round the expected frequencies.

$$\begin{aligned}
 \text{Expected frequency (cell A and B)} &= E(A \text{ and } B) \\
 &= n \left( \frac{\text{Row Total}}{n} \cdot \frac{\text{Column Total}}{n} \right) = \frac{\text{Row Total} \cdot \text{Column Total}}{n}
 \end{aligned}$$

If the variables are independent, the expected frequencies and the observed frequencies should be the same.

The test statistic here will involve looking at the difference between the expected frequency and the observed frequency for each cell. Then you want to find the “total difference” of all of these differences. The larger the total, the smaller the chances that you could find that test statistic given that the assumption of independence is true. That means that the assumption of independence is not true.

How do you find the test statistic? First, compute the differences between the observed and expected frequencies. Because some of these differences will be positive and some will be negative, you need to square these differences. These squares could be large just because the frequencies are large, so you need to divide by the expected frequencies to scale them. Then finally add up all of these fractional values. This process finds the variance, and we use a chi-square distribution to find the critical value or p-value. Hence, sometimes this test is called a chi-square test.

The  $\chi^2$ -test is a statistical test for testing the independence between two variables. It can be used when the data are obtained from a random sample, and when the expected value ( $E$ ) from each cell is 5 or more.

The formula for the  $\chi^2$ -test statistic is:  $\chi^2 = \sum \frac{(O-E)^2}{E}$ .

Use  $\chi^2$ -distribution with degrees of freedom

$df = (\text{the number of rows} - 1)(\text{the number of columns} - 1)$ , that is,  $df = (R - 1)(C - 1)$ .

where  $O$  = the observed frequency (sample results) and

$E$  = the expected frequency (based on  $H_0$  and the sample size).

Is there a relationship between autism spectrum disorder (ASD) and breastfeeding? To determine if there is, a researcher asked mothers of ASD and non-ASD children to say what time-period they breastfed their children. Does the data provide enough evidence to show that breastfeeding and ASD are independent? Test at the 1% level.

	Length of Breastfeeding				
ASD	None	Less than 2 months	2 to 6 months	Over 6 months	Total
Yes	241	198	164	215	818
No	20	25	27	44	116
Total	261	223	191	259	934

(Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006.)

### Solution

The question is asking if breastfeeding and ASD are independent. The correct hypothesis is:

$H_0$  : Autism spectrum disorder and length of breastfeeding are independent.

$H_1$  : Autism spectrum disorder and length of breastfeeding are dependent.

There are 2 rows and 4 columns of data. We must compute the Expected count for each of the  $2 \times 4 = 8$  cells.

The expected counts for each cell are found by the formula:

$$\text{Expected Value} = \frac{\text{Row Total} \cdot \text{Column Total}}{\text{Grand Total}}$$

It will be helpful to make a table for the expected counts and another one for each of the  $\frac{(O-E)^2}{E}$  values to aid in computing the test statistic.

Observed	ASD	None	Below 2	2-6	Over 6	Total
	Yes	241	198	164	215	818
	No	20	25	27	44	116
	Total	261	223	191	259	934

Expected	ASD	None	Below 2	2-6	Over 6	Total
	Yes	$\frac{818 \cdot 261}{934} = 228.585$	$\frac{818 \cdot 223}{934} = 195.304$	$\frac{818 \cdot 191}{934} = 167.278$	$\frac{818 \cdot 259}{934} = 226.833$	818
	No	$\frac{116 \cdot 261}{934} = 32.415$	$\frac{116 \cdot 223}{934} = 27.696$	$\frac{116 \cdot 191}{934} = 23.722$	$\frac{116 \cdot 259}{934} = 32.167$	116
	Total	261	223	191	259	934

$\frac{(O-E)^2}{E}$	ASD	None	Below 2	2-6	Over 6	Total
	Yes	$\frac{(241-228.585)^2}{228.585} = 0.6743$	$\frac{(198-195.304)^2}{195.304} = 0.0372$	$\frac{(164-167.278)^2}{167.278} = 0.0642$	$\frac{(215-226.833)^2}{226.833} = 0.6173$	
	No	$\frac{(20-32.415)^2}{32.415} = 4.7552$	$\frac{(25-27.696)^2}{27.696} = 0.2624$	$\frac{(27-23.722)^2}{23.722} = 0.4531$	$\frac{(44-32.167)^2}{32.167} = 4.3529$	11.217

The test statistic is the sum of all eight  $\frac{(O-E)^2}{E}$  values:  $\chi^2 = \sum \frac{(O-E)^2}{E} = 11.217$ .

The critical value for a right-tailed  $\chi^2$ -test with degrees of freedom  $df = (R-1)(C-1) = (2-1)(4-1) = 3$  is found using a  $\chi^2$ -distribution  $\alpha = 0.01$  right-tail area. The critical value is  $\chi^2 = \text{CHISQ.INV.RT}(0.01,3) = 11.3449$ . See Figure 10-5.

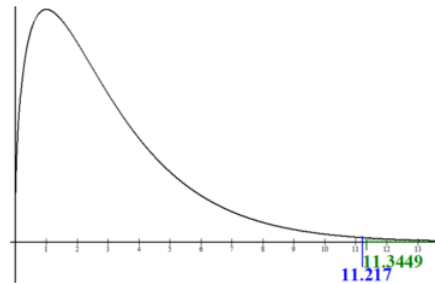
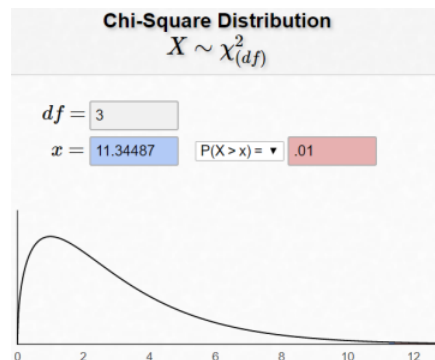


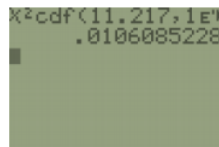
Figure 10-5: Values of  $\chi^2$  and  $\chi^2_{\alpha}$  on a  $\chi^2$ -test with  $df = 3$  and  $\alpha = 0.01$ .

Alternatively, use the online calculator: <https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html>.



Since the test statistic  $\chi^2 = 11.217$  is not in the rejection area, our decision is to fail to reject  $H_0$ .

There is not enough evidence to show a relationship between autism spectrum disorder and breastfeeding.



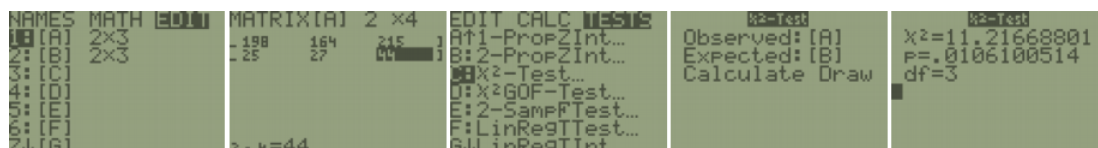
If we were asked to find the p-value, you would just find the area to right of the test statistic (always a right-tailed test) using your calculator or Excel. This gives a p-value = 0.0106, which is more than  $\alpha = 0.01$ ; therefore, we do not reject  $H_0$ .

You can also use the  $\chi^2$ -Test shortcut keys on your calculator to get a p-value, see directions below.

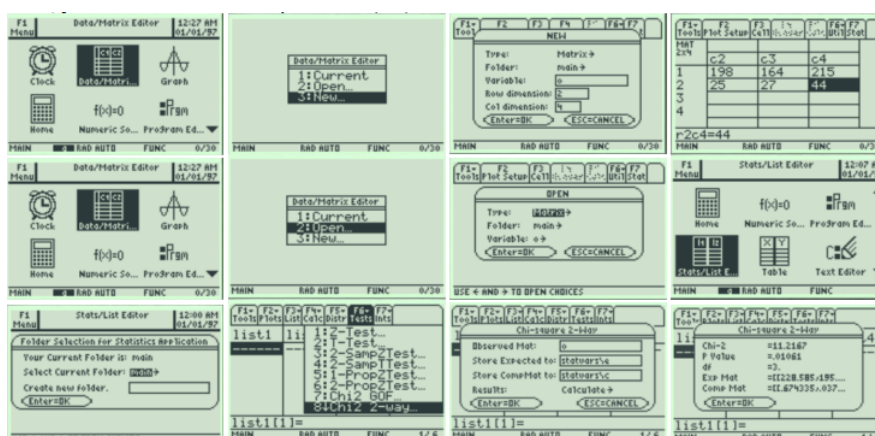
**TI-84:** Press the [2<sup>nd</sup>] then [MATRX] key. Arrow over to the EDIT menu and 1:[A] should be highlighted, press the [ENTER] key. For a  $m \times n$  contingency table, type in the number of rows ( $m$ ) and the number of columns ( $n$ ) at the top of the screen so that it looks like this: MATRIX[A]  $m \times n$ . For a  $2 \times 4$  contingency table, the top of the screen would look like this: MATRIX[A]  $2 \times 4$ . As you hit [ENTER], the table will automatically widen to the size you put in. Now enter all of the observed values in their proper positions. Then press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [C:  $\chi^2$ -Test] and press the [ENTER] key. Leave the default as Observed:[A] and Expected:[B], arrow down to [Calculate] and press the [ENTER] key.

The calculator returns the  $\chi^2$ -test statistic and the p-value.

If you go back to the matrix menu [2<sup>nd</sup>] then [MATRX] key, arrow over to EDIT and choose 2:[B], you will see all of the expected values.



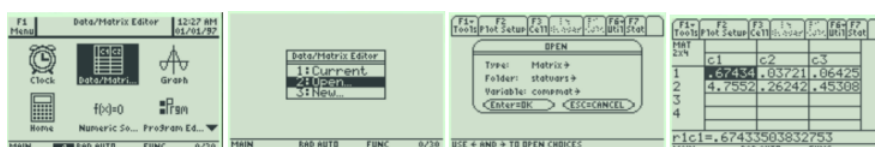
**TI-89:** First you need to create the matrix for the observed values: Press [Home] to return to the Home screen, press [Apps] and select Data/Matrix Editor. A menu is displayed, select 3:New. The New dialog box is displayed. Press the right arrow key to highlight 2:Matrix, and press [ENTER] to choose Matrix type. Press the down arrow key to highlight 1:Main, and press [ENTER], to choose main folder. Press the down arrow key, and then enter the letter *o* for the name in the Variable field. Enter 2 for Row dimension and 4 for Column dimension. Press [ENTER] to display the matrix editor. Enter the observed value (do not include total row or column). Important: Next time you use this test instead of option 3:New, choose 2: Open. The open dialog box is displayed. Press the right arrow key to highlight 2:Matrix, and press [ENTER] to choose Matrix type. Press the down arrow key to make sure you are in the Main folder and that your variable says *o*. Press [Apps], and then select Stats/List Editor. To display the Chi-square 2-Way dialog box, press 2<sup>nd</sup> then F6 [Tests], then select 8: Chi-2 2-way. Enter in the Observed Mat: *o*; leave the other rows alone: Store Expected to: statvars\e; Store CompMat to: statvars\c. This will store the expected values in the matrix folder statvars with the name expmat, and the  $(o - e)^2/e$  values in the matrix compmat. Press the [ENTER] key to calculate. The calculator returns the  $\chi^2$ -test statistic and the p-value. If you go back to the matrix menu, you will see some of the expected and  $(o - e)^2/e$  values.



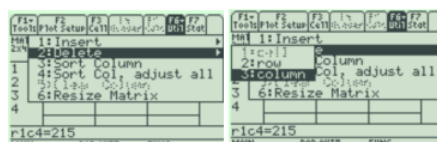
To see all the expected values, select [APPS] and select Data/Matrix Editor. Select 2:Open, change the Type to Matrix, change the Folder to statvars, and change the Variable to expmat.



To see all the  $(o - e)^2/e$  values, select [APPS] and select Data/Matrix Editor. Select 2:Open, change the Type to Matrix, change the Folder to statvars, and change the Variable to compmat.



If you need to delete a row or column, move the cursor to the row or column that you want to delete, then select F6 Util, then 2:Delete, then choose row or column, then enter. To add a row or column, just arrow over to the new row or column and type in the observed values.



The sample data below show the number of companies providing dental insurance for small, medium and large companies. Test to see if there is a relationship between dental insurance coverage and company size. Use  $\alpha = 0.05$ .

	Size of the Company		
Dental Insurance	Small	Medium	Large
Yes	21	25	19
No	46	39	10

### Solution

State the hypotheses.

$H_0$  : Dental insurance coverage and company size are independent.

$H_1$  : Dental insurance coverage and company size are dependent.

Compute the expected values by taking each row total times column total, divided by grand total.

$O = \text{observed}$	Size of the Company			
Dental Insurance	Small	Medium	Large	Total
Yes	21	25	19	65
No	46	39	10	95
Total	67	64	29	160

For the small companies with dental insurance:  $(65 \cdot 67)/160 = 27.21875$

small companies without dental insurance:  $(95 \cdot 67)/160 = 39.78125$

medium companies with dental insurance:  $(65 \cdot 64)/160 = 26$ , etc. See table below.

$E = \text{expected}$	Size of the Company			
Dental Insurance	Small	Medium	Large	Total
Yes	27.21875	26	11.78125	65
No	39.78125	38	17.21875	95
Total	67	64	29	160

Compute the test statistic.

$\frac{(O - E)^2}{E}$	Size of the Company		
Dental Insurance	Small	Medium	Large
Yes	$\frac{(21 - 27.21875)^2}{27.21875} = 1.42082$	$\frac{(25 - 26)^2}{26} = 0.03846$	$\frac{(19 - 11.78125)^2}{11.78125} = 4.42316$
No	$\frac{(46 - 39.78125)^2}{39.78125} = 0.97214$	$\frac{(39 - 38)^2}{38} = 0.02632$	$\frac{(10 - 17.21875)^2}{17.21875} = 3.02637$

Test statistic is  $\chi^2 = \sum \frac{(O - E)^2}{E} = 1.42082 + 0.03846 + 4.42316 + 0.97214 + 0.02632 + 3.02637 = 9.9073$

Use technology to find the p-value using the chi-square cdf with  $df = (R - 1)(C - 1) = (2 - 1)(3 - 1) = 2$ .

$\chi^2 = 9.907263904$
$P = .007057729$
$df = 2$

Using the TI-Calculator, we find the p-value = 0.0071.

The p-value is less than  $\alpha$ ; therefore, reject  $H_0$ .

There is enough evidence to support the claim that there is a relationship between dental insurance coverage and company size.

This page titled [10.3: Test for Independence](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Rachel Webb](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.