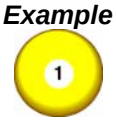
Chapter 7: Cross Table Analysis

Chi-Squared Test of a Contingency Table

7.1 Chi-Squared Test of a Contingency Table

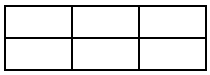
The chi-squared test is used to determine if there is enough evidence to infer that two are related and to infer that differences exist among two qualitative variables. Completing both objectives entails to two different criteria. The following is an Example to see how this is done.

Example



Example (1)

One of the issues that came up in a recent national election in many future elections) is how to deal with a sluggish should governments cut spending, raise taxes, inflate the more money), or do none of the above and let the deficit rise politicians need to know which parts of the electorate suppose that a random sample of 1,000 people was asked which and their political affiliations. The possible responses to the affiliation were Democrat, Republican, and Independent. The responses were summarized in cross-classification table, shown below. Do this conclude that political affiliation affects support for the ecology.



	Political Aff.	
Economic Opinions	Democrat	Republican
Cut spending	101	282
Raise taxes	38	67
Inflate the economy	131	88
Let deficit increase	61	90

Solution

Solution:

One way to solve the problem is to consider the contingency table. The variables are economic affiliation. Both are qualitative. The values of economic "raise taxes," "inflate the economy," and "let deficit increase political affiliation are "Democrat," "Republican" and "Independent" objective is to analyze the relationship between the two variables. Specifically, we want to know whether one variable affects the other.

Another way of addressing the problem is to determine whether differences exist among Democrats, Republicans, and Independents. In other words, we treat each political group as a separate population. Each population has four possible values, represented by the four economic options. (We can also answer the question by treating the economic options as populations and the political affiliations as the values of the random variable.) Here the problem objective is to compare three populations.

As you will shortly discover, *both objectives lead to the same test. Consequently, we can address both objectives at the same time.*

The null hypothesis will specify that there is no relationship between the two variables. We state this in the following way.

H_o: The two variables are independent.

The alternative hypothesis specifies that one variable affects the other, which is expressed as

H_A: The two variables are dependent.

If the null hypothesis is true, political affiliation and economic option are independent of one another. *This means that whether someone is a Democrat, Republican, or Independent does not affect his economic choice. Consequently, there is no difference among Democrats, Republicans, and Independents in their support for the four economic options. If the alternative hypothesis is true, political affiliation does affect which economic option is preferred. Thus, there are differences d is likely to among the three political groups.*

The test statistic is

$$x^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Where k is the number of cells in the contingency table. *The null hypothesis for the chi-squared test of a contingency table only states that the two variables are independent. However, we need the probabilities in order to compute the expected values (e_j), which in turn permits us to calculate the value of the test statistic. (The entries in the contingency table are the observed values, o_i. The question immediately arises: from where do we get the probabilities? The answer is that they will come from the data after we assume that the null hypothesis is true.*

If we consider each political affiliation to be a separate population, each column of the contingency table represents an experiment with four cells. If the null hypothesis is true, the three experiments should produce similar proportions in each cell. We can estimate the cell probabilities by calculating the total in each row and dividing by the sample size. Thus,

$$\text{P(cut spending)} \approx \frac{444}{1{,}000}$$

$$\text{P(raise taxes)} \approx \frac{250}{1{,}000}$$

$$\text{P(let deficit increase)} \approx \frac{176}{1{,}000}$$

We can calculate the expected values for each cell in the three by multiplying these probabilities by the total number of political group. By adding down each column, we find that there are residents who identified themselves as Democrats (331), 527 as Republicans and 142 as independents.

Expected Values of the Economic Options of Democrats

EONOMIC OPTION	EXPECTED VALUE
Cut spending	$331 \times \frac{444}{1{,}000} = 146.96$
Raise Taxes	$331 \times \frac{130}{1{,}000} = 143.03$
Inflate economy	$331 \times \frac{250}{1{,}000} = 82.75$
Let deficit increase	$331 \times \frac{176}{1{,}000} = 58.26$

Expected Values of the Economic Options of Republicans

EONOMIC OPTION	EXPECTED VALUE
Cut spending	$527 \times \frac{444}{1{,}000} = 233.99$
Raise Taxes	$527 \times \frac{130}{1{,}000} = 68.51$
Inflate economy	$527 \times \frac{250}{1{,}000} = 131.75$
Let deficit increase	$527 \times \frac{176}{1{,}000} = 92.75$

Expected Values of the Economic Options of Independents

EONOMIC OPTION	EXPECTED VALUE
Cut spending	$142 \times \frac{444}{1{,}000} = 63.05$
Raise taxes	$142 \times \frac{130}{1{,}000} = 18.46$
Inflate economy	$142 \times \frac{250}{1{,}000} = 35.50$
Let deficit increase	$142 \times \frac{176}{1{,}000} = 24.99$

Notice that the expected values are computed by multiplying the column total by the row total and dividing by the sample size.

Expected Frequencies for a Contingency Table

7.2 Expected Frequencies for a Contingency Table

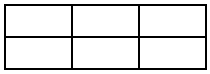
The expected frequency of the cell in column j and row i is

$$e_{ij} = \frac{(\text{Column j total})(\text{Row i total})}{\text{Sample size}}$$

The expected cell frequencies are shown in parentheses in the Table below, the expected cell frequencies should satisfy the rule of five.

Contingency Table of Example 3

ECONOMIC OPTIONS	POLITICAL AFFILIATION		
	DEMOCRATE	REPUBLIC	INDEPENDENT
Cut spending	101 (146.96)	282 (233.99)	61 (63.05)
Raise Taxes	38 (43.03)	67 (68.51)	25 (18.46)
Inflate economy	131 (82.75)	88 (131.75)	31 (35.50)
Let deficit increase	61 (58.26)	90 (92.75)	25 (24.99)



We can now calculate the value of the test statistic. It is

$$x^2 = \sum_{i=1}^{12} \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(101-146.96)^2}{146.96} + \frac{(38-43.03)^2}{43.03} + \frac{(131-82.75)^2}{82.75} + \frac{(61-58.26)^2}{58.26}$$
$$+ \frac{(282-233.99)^2}{233.99} + \frac{(67-68.51)^2}{68.51} + \frac{(88-131.75)^2}{131.75} + \frac{(90-92.75)^2}{90.75}$$
$$+ \frac{(61-63.05)^2}{63.05} + \frac{(25-18.46)^2}{18.46} + \frac{(31-35.50)^2}{35.50} + \frac{(25-24.99)^2}{24.99}$$
$$= 70.675$$

Notice that we continue to use a single subscript in the formula of the test statistic when we should use two subscripts, one for the rows and one for the columns. *We feel that it is clear that for each cell, we need to calculate the squared difference between the observed and expected frequencies divided by the expected frequency*. We don't believe that the satisfaction of using the mathematically correct notation would overcome the unnecessary complication.

Rejection Region

Rejection Region

To determine the rejection region, we need to know the number of degrees of freedom associated with this x2 - statistic. The number of degrees of freedom for a contingency Table with r rows and c columns is

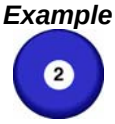
For Example 3, the number of degrees of freedom is

$$\text{d.f.} = (r - l)(c - 1) = (4 - 1)(3 - 1) = 6$$

If we use a 5% significance level, the rejection region is $x^2 > x^2_\alpha, x^2.05,6 = 12.5916$

Because x^2 = 70.675, we reject the null hypothesis and conclude that evidence of a relationship between political affiliation and support for nomic options. It follows that the three political affiliations differ in their for the four economic options. We can see from the data that Republicans favor cutting spending, whereas Democrats prefer to inflate the economy.

Example



Example (2)

The operations manager of a company that manufactures shirts whether there are differences in the quality of workmanship am shifts. She randomly selects 600 recently made shirts and scarf. Each shirt is classified as either perfect or flawed, and the shift also recorded. The accompanying Table summarizes the number into each cell. Do these data provide sufficient evidence at the 5 to infer that there are differences in quality among the three?

Contingency Table Classifying Shirts

		SHIFT
SHIFT CONDITION	1	2
Perfect	240	191
Flawed	10	9

Solution

Solution:

The problem objective is to compare three populations (the shirt three shifts). The data are qualitative because each shirt will be perfect or flawed. This problem - objective / data - type combination statistical procedure to be employed is the chi-squared test of a. The null and alternative hypotheses are as follows.

H_o: The two variables are independent.

H_A: The two variables are dependent.

Test statistics:

$$x^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \qquad \text{d.f.} = (r - 1)(c - 1)$$

We calculated the row and column totals and used them to determine the expected values. For example, the expected number of perfect shirts produced in shift 1 is

$$e_1 = \frac{250 \times 570}{600} = 237.5$$

The remaining expected values are computed in a like manner. The original Table and expected values are shown in the Table below.

SHIRT CONDITION	SHIFT 1	2	3	TOTAL
Perfect	240 (237.5)	191 (190.0)	139 (142.5)	570
Flawed	10 (12.5)	9 (10.0)	11 (7.5)	30
TOTAL	250	200	150	600

The value of the test statistic is

$$x^2 = \sum_{i=1}^{6} = \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(240 - 237.5)^2}{237.5} + \frac{(10 - 12.5)^2}{12.5} + \frac{(191 - 190.0)^2}{190.0} + \frac{(9 - 10.0)^2}{10.0}$$

$$+ \frac{(139 - 142.5)^2}{142.5} + \frac{(11 - 7.5)^2}{7.5}$$

$$= 2.36$$

Conclusion: Do not reject the null hypothesis

We can measure how strong is the relationship between the two variables using (sort of a correlation coefficient called contingency coefficient CCC)

$$CC = \sqrt{\frac{x^2}{x^2 + n}}$$