Chapter 5.6: Tests for Independence

Previously, we used parametric tests, e.g. is there any evidence that $p < 0.5$?

Now we want to consider a nonparametric test for evidence of a relationship between two variables.

Example

The table contains data from the 1991 US general social survey of of level of confidence in the TV press and average hours of daily tv watching.

Is there any evidence of a relationship between confidence in the press and level of tv viewing?

As far as the people running the press, you would have …	Average hours of daily tv watching			
	0-1 hours	2-4 hours	5 or more	Total
A good deal of confidence	276	41	17	334
Only some confidence	196	174	47	417
Hardly any confidence	130	97	15	242
Total	602	312	79	993

Independence of variables

We have two categorical variables:

X = confidence in the press

Y = level of tv viewing

X and Y are *independent* if $P(X = x, Y = y) = P(X = x)\, P(Y = y)$ for every possible value of x and y.

Formulation as a hypothesis test

Our experimental hypothesis is that there is a relationship between X and Y, that is that they are not independent.

H_0: X and Y are independent

H_1: X and Y are not independent

Now we proceed like any hypothesis test. Assume H_0 is true and try to see if the data provide evidence against this assumption.

Estimating the marginal distributions

What numbers would we expect to see in each cell if the variables really were independent?

As far as the people running the press, you would have …	Average hours of daily tv watching			
	0-1 hours	2-4 hours	5 or more	Total
A good deal of confidence	276	41	17	334
Only some confidence	196	174	47	417
Hardly any confidence	130	97	15	242
Total	602	312	79	993

We can start by estimating the marginal distributions by the marginal frequencies.

602/993 = 0,60624

As far as the people running the press, you would have …	Average hours of daily tv watching			
	0-1 hours	2-4 hours	5 or more	Total
A good deal of confidence				0,34
Only some confidence				0,42
Hardly any confidence				0,24
Total	0,60624	0,3142	0,07956	1

Estimating the joint distribution

Now, assuming independence, we can estimate P(X = x, Y = y) by the product of the estimated marginal distributions.

As far as the people running the press, you would have …	Average hours of daily tv watching			
	0-1 hours	2-4 hours	5 or more	Total
A good deal of confidence	0,20391	0,10568	0,02676	0,34
Only some confidence	0,25459	0,13194	0,03341	0,42
Hardly any confidence	0,14775	0,07657	0,01939	0,24
Total	0,60624	0,3142	0,07956	1

0,20391 = 0,34 x 0,60624

Calculating expected values

We know that our sample has 993 people in total. Therefore multiply the estimated probabilities in the last table by 993 to get expected values.

As far as the people running the press, you would have …	Average hours of daily tv watching 0-1 hours	2-4 hours	5 or more	Total
A good deal of confidence	202,485	104,943	26,572	334
Only some confidence	252,804	131,021	33,1752	417
Hardly any confidence	146,711	76,0363	19,2528	242
Total	602	312	79	993

202,485 = 0,20391 x 993

A more direct way: 202,485 = 334 x 602 / 993

A general formula is:

Expected value in cell i,j = total in row i x total in row j / sample size

The test statistic

If the two variables really are independent, we would expect the observed and expected values to be similar. To measure this we calculate the test statistic:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

As far as the people running the press, you would have …	Average hours of daily tv watching 0-1 hours	2-4 hours	5 or more	
A good deal of confidence	26,6903	38,9609	3,44811	
Only some confidence	12,7635	14,0983	5,76106	
Hardly any confidence	1,90345	5,77986	0,9394	
				110,34

$(276 – 202{,}485)^2 / 202{,}485 + … + (15 – 19{,}2528)^2 / 19{,}2528 = 110{,}34$

The chi squared distribution

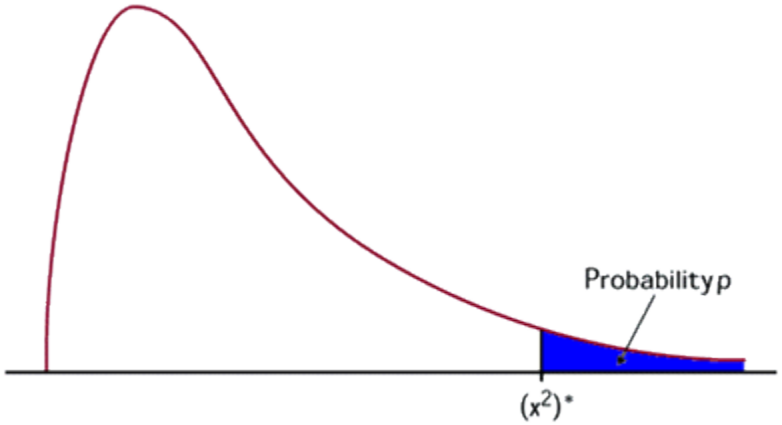
If the two variables really are independent, it is known that the test statistic is generated from a *chi-squared distribution* with:

degrees of freedom = (number of rows – 1) x (number of columns -1)

In our case, we have 3 rows and 3 columns so the degrees of freedom are (3 – 1) x (3 – 1) = 4.

Calculating the p value

Large values of the test statistic mean that observed and expected numbers are different. Therefore we should decide to reject the null hypothesis if the number is too high. We can calculate the p-value as below.



In our case, we have p = 6,14E-23, almost zero.

Finishing the test

As earlier, if we fix a significance level, α = 0,05 for example, we can compare the p value with α to conclude the test.

At a 5% significance level, we would reject the hypothesis of independence between the opinion about the press and time spent watching tv.

There is strong evidence of a relationship between the two variables.

Computation in Excel

Assume the observed frequencies are in cells B3:D5.

276	41	17
196	174	47
130	97	15

Assume the expected frequencies are in cells B10:D12.

202,485	104,943	26,572
252,804	131,021	33,1752
146,711	76,0363	19,2528

6,14E-23 = PRUEBA.CHI(B3:D5;B10:D12)

A small problem

The chi-squared test is only reliable if all expected frequencies are > 1 and at least 80% of expected frequencies are > 5.

If this is not the case, we may have to combine rows (or columns) to provide accurate results.

Example

The following data are the number of votes emitted by undergraduate students in the different campuses of the UC3M in favour of each of the rectoral candidates in one of the previous university elections:

	Luciano Parejo	Francisco Marcellán	Daniel Peña
Getafe	954	525	330
Leganes	130	534	187
Colmenarejo	665	21	14

Is there any evidence of a relationship between campus and voting intention of Carlos III students?

Example

The following data (reported by Paul Gingrich) come from a 1988 survey of adults in Newfoundland, Canada:

Respondents who knew someone on social assistance, were more likely to feel that welfare rates were too low, ...

	Knows Someone on Social Assistance		Row
Welfare Spending	Yes	No	Totals
Too Little	40	6	46
About Right	16	13	29
Too Much	9	7	16
Column Totals	65	26	91

Is there any evidence of a relationship between opinion on welfare spending and knowing people on social assistance?

Example

The following data (reported by Paul Gingrich) come from a survey of adults in Edmonton, Canada on opinions about whether the trades unions are responsible for unemployment.

	Political Preference	
Opinion	PC	Liberal
1	9	3
2	7	5
3	7	11
4	28	3
5	51	12
6	54	7
7	58	12

Is there any evidence of a relationship between opinion about the trades unions causing unemployment and political preference?