Big Data Analytics

Lab 1 EXTRA A – Chi-squared test

UNIVERSITÄT LUZERN

Chi2

Let's assume the following contingency table

	LEFT	RIGHT	*Total*
British	3	23	*26*
Italian	32	5	*37*
Total	*35*	*28*	*63*

We want to evaluate how likely it is that any observed difference between the sets arose by chance. For doing that, let's employ the Pearson's chi-squared test (Chi2)

$$X_c^2 = \sum \frac{(O - E)^2}{E}$$

where: c = degrees of freedom; O = observed frequency; E = expected frequency

Chi2

What do we mean by expected frequency?

To calculate the expected frequency for each cell of the table we have first to consider the *null hypothesis*, which in this case is that the numbers in each cell are proportionately the same in the British sample as they are in the Italian sample

We therefore construct a parallel table in which the proportions are exactly the same for both samples

How to do it?

Chi2

	LEFT	RIGHT	*Total*
British	3	23	*26*
Italian	32	5	*37*
Total	*35*	*28*	*63*

The proportions are obtained from the totals column in the previous table and are applied to the totals row

	E left	E right	(O-E) for E left	(O-E) for E right	(O-E)^2/E for E left	(O-E)^2/E for E right
British	14.44	11.56				
Italian	20.55	16.44				

For instance, in table above, in column (E left) (26/63) x 35=14.44; (37/63) x 35=20.55; in column (E right) (26/63) x 28 = 11.55; (37/63) x 28 = 16.44

Chi2

	LEFT	RIGHT	*Total*
British	3	23	*26*
Italian	32	5	*37*
Total	*35*	*28*	*63*

	E left	E right	(O-E) for E left	(O-E) for E right	(O-E)^2/E for E left	(O-E)^2/E for E right
British	14.44	11.56	-11.44	11.44	9.06	11.33
Italian	20.55	16.44	11.44	-11.4444	6.37	7.96
Total					*15.43*	*19.29*

Here the χ^2 is: (15.43+19.29)=34.74

Clearly, the larger the difference between the observations and the expectations (O − E in the equation), the bigger the chi-square will be

To decide whether the difference is big enough to be statistically significant, you compare the chi-square value to a critical value (after having identified the related degree of freedom…)

Chi2

	LEFT	RIGHT	*Total*
British	3	23	*26*
Italian	32	5	*37*
Total	*35*	*28*	*63*

	E left	E right	(O-E) for E left	(O-E) for E right	(O-E)^2/E for E left	(O-E)^2/E for E right
British	14.44	11.56	-11.44	11.44	9.06	11.33
Italian	20.55	16.44	11.44	-11.4444	6.37	7.96
Total					*15.43*	*19.29*

Here the degree of freedom is 1 (i.e., (# of columns minus 1) x (# of rows minus 1) (not counting the row and column containing the totals)

If we now look at a table of χ^2 distribution the probability attached to the χ^2 with 1 degree of freedom is, we find a p-value <0.001 given our 34.74 value above (i.e., we can reject the null hyp. of no relationship in a pretty confident way…)

Chi2

The `textstat_keyness` command within Quanteda does a very similar exercise

It considers: 1) in the 2 rows the target vs. the reference text; 2) in the first column the frequency of the feature we are interested about (i.e., say “American”) as it appears in the two set of texts from the `DfM`; 3) in the second column the frequency of all the other features in the two set of texts

It also implements, by default, a Yates correction. Basically it subtracts 0.5 from the numerator of the χ^2 formula

This aims at correcting the error introduced by assuming (as we do with chi2) that the discrete probabilities of frequencies in the table can be approximated by a continuous (chi-squared) distribution

Chi2

Finally, remember that chi2 is a *non-parametric test*

Parametric tests use data from a sample to draw conclusions about a population, and the parameters of that population are expected to meet certain assumptions

Non-parametric tests do not require assumptions about the underlying population and do not test hypotheses about population parameters

Categorical data, and data that are not normally distributed, can be analyzed with non-parametric statistics

After all, with categorical variables, we can't calculate a mean or standard deviation. Instead, we have just frequencies