

Research Methods

Chi² (X²)

Chi²

- Non-parametric test
- Not based on any assumption or distribution of variables
- There are a number of tests based on the Chi² statistic
- Two main types of test:
 - Chi² for goodness of fit
 - Chi² for independence

Chi²

- Can come from experimental, questionnaire, or observational research
- Participants are categorised in some manner
- Number of people in each category are counted
- Data takes the form of frequencies of occurrences
- Despite being distribution free, the sample size should be above 20

Chi²

- Used to investigate whether distributions of **categorical variables** differ from one another
 - Nominal data (frequencies by categories)
- Numbers only represent the category of which the participant is a member
 - The count of total people in that category
- "Not based on any assumption or distribution of variables"...
- But - follows a specific distribution known as *Chi² distribution*

Chi²

- The chi² measures the difference between what is observed (O) and what is expected (E) (according to an assumed hypothesis)
- Chi² statistic (X^2) is computed to measure the amount of discrepancy between E and O
- A large discrepancy = large X^2 value
- Indicates the data do not fit the null hypothesis and the null hypothesis should be rejected
 - Research hypothesis therefore gets accepted

Chi²

- For results of a Chi-Square test to be reliable:
 - Data must be from a random sample from the population
- Problems arise when the expected frequencies are very small.
 - As a rule of thumb, Chi Square should not be used if more than 20% of the expected frequencies have a value of less than 5 (it does not matter what the observed frequencies are).
 - You can get around this in two ways:
 - either combine some categories (if meaningful, in your experiment), or obtain more data (make the sample size bigger).

Chi² for independence

- Enables us to explain whether or not two attributes are associated
- Let's have a look at a study example....
- Suppose we suspect rural Welsh people were more likely to vote to remain in the EU, and urban Welsh people were more likely to vote to leave the EU. In this case we are looking for a relationship between where people live and how they vote.
 - We will use a fictitious pole conducted on 100 people asking where they live and how they voted.

7

Calculating Chi²

1. Create a chi-square **contingency** table
2. Note observed frequencies per cell
3. Note observed frequencies in each class of one event (row-wise) and number in each group of the other event (column-wise)
4. Determine expected frequency (E) in each group of sample/cell
5. Calculate chi² statistic
6. Work out degrees of freedom (DF)
7. Determine whether significant difference or not using critical values table

8

1. Create a contingency table

- Each row will be how they voted, giving us two columns for remain and leave
- Each row will be where they live, giving us two rows for rural and urban

	Remain	Leave	Row frequency
Rural			
Urban			
Column frequency			

9

2. Note observed frequencies (O) per cell

	Remain	Leave	Row frequency
Rural	45	15	
Urban	10	30	
Column frequency			N = 100

3. Note row and column frequencies

	Remain	Leave	Row frequency
Rural	45	15	60
Urban	10	30	40
Column frequency	55	45	N = 100

4. Determine expected frequencies (F_e)

- Expected frequencies are the frequencies we expect in our sample if the null hypothesis holds.
- If how you vote and where you live are independent in our population, then we expect this in our sample too.
 - No association – i.e., where you live makes no difference to how you vote.
- This is done using the marginal totals and overall total to compute expected counts for each cell of the table.
- To find the expected count for each cell in the table we multiply the marginal row and column totals for that cell and divide by the overall total.

4. Determine expected frequencies (F_e)

- Formulaically for each cell this is:

$$E = \frac{\text{row total} \times \text{column total}}{\text{Sample size}}$$

	Remain	Leave	Row frequency
Rural	45	15	60
Urban	10	30	40
Column frequency	55	45	N = 100

$$F_e = \frac{(\text{row total})(\text{column total})}{N}$$

13

4. Determine expected frequencies

	Remain	Leave	Row frequency
Rural	$E = 60 \times 55 / 100 =$ $F_e = 33$	$E = 60 \times 45 / 100 =$ $F_e = 27$	60
Urban	$E = 40 \times 55 / 100 =$ $F_e = 22$	$E = 40 \times 45 / 100 =$ $F_e = 18$	40
Column frequency	55	45	N = 100

14

5. Calculate chi² statistic

- The statistical question becomes:
 - Are the observed counts so different from the expected counts that we can conclude a relationship between the two variables?
- To conduct this test we compute a Chi-square test statistic where we compare each cell's observed count to it's respective expected count
- This Chi-square test statistic is calculated as follows:

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

15

5. Calculate chi² statistic

$$X^2 = \sum \frac{(o - e)^2}{e}$$

Explanation of symbols:

X² is the chi square

Σ (Greek letter pronounced sigma) means "Add up the results of each of what follows"

O is the observed values

E is the expected values

The small 2 (superscript) means we need to square the value – i.e., multiply it by itself

5. Calculate chi² statistic

Observed	Expected
45	33
15	27
10	22
30	18

$$X^2 = \frac{(45 - 33)^2}{33} + \frac{(15 - 27)^2}{27} + \frac{(10 - 22)^2}{22} + \frac{(30 - 18)^2}{18}$$

Work out what is in the brackets first

$$X^2 = \frac{12^2}{33} + \frac{-12^2}{27} + \frac{-12^2}{22} + \frac{12^2}{18}$$

Then calculate the square

$$X^2 = \frac{144}{33} + \frac{144}{27} + \frac{144}{22} + \frac{144}{18}$$

Then do the division

$$X^2 = 4.364 + 5.333 + 6.545 + 8 = 24.24$$

(present to two decimal places)

6. Work out df

DF is a common term in statistical analysis having to do with the number of individual data points that are free to vary given that overall summary values are known

The degrees of freedom is basically a number that determines the exact shape of our distribution. It is calculated as

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

$$df = (2-1) (2-1) = 1 \times 1 = 1$$

7. Determine whether significance from critical values table

To see if there is a significant association we need to look at the critical values table.

If the chi-square calculated value is greater than the chi-square critical value, then you reject your null hypothesis (and accept your research hypothesis). If smaller – no significant association

Critical values of the Chi-square distribution with d degrees of freedom

Probability of exceeding the critical value							
d	0.05	0.01	0.001	d	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

19

7. Determine whether significance from critical values table

DF = 1

Chi² = 24.24

Sig. value = 0.05

The critical value in the chi-square table for df of 1 and 0.05 sig. is 3.841

Our value of 24.24 is greater than this value, so our results are significant

This means we can say there is a significant link between where people live and Brexit vote; people have different views on Brexit and these views are linked to where they live.

Critical values of the Chi-square distribution with d degrees of freedom

Probability of exceeding the critical value							
d	0.05	0.01	0.001	d	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

21

Calculating effect size of Chi-square analysis

- For chi-square analysis, the effect size can be calculated using Cramer's phi Φ

$$\Phi = \sqrt{\frac{x^2}{(N)}}$$

$$\Phi = \sqrt{24.24/100}$$

$$\Phi = \sqrt{0.242} = 0.49$$

Note: N is number of participants in total. Also, the smaller of any dfs is included when a category has more than two options

Cohen (1988) produced effect size conventions for phi – see table in your textbooks, which takes into account df

Chi-square Write-up

- E.g.,
- Differences in Brexit vote, depending on where participants lived was found in this study. Seventy-five percent of participants living in rural areas of Wales voted to remain in the EU (45/60). While only 25% of participants living in urban areas of Wales voted to remain in the EU (10/40). A chi-square analysis of the difference between remain and leave votes across the different areas was significant, $\chi^2 = (1, N = 100) = 24.24, p < 0.05$. Phi shows that the effect size of the association was large (0.49).
- (notice how the letters are in italics and the first percentage to start the sentence is in word format)

And now its your turn!

- 'Dummy data' is available on Moodle
- You need to work through the 7 steps to determine if males and females have different preferences for pets.
