

Research Methods I

Chi² (X^2)

Stacey Coleman

Chi²

- Non-parametric test
- Not based on any assumption or distribution of variables
- Two types of test:
 - *Chi² for goodness of fit*
 - *Chi² for independence*

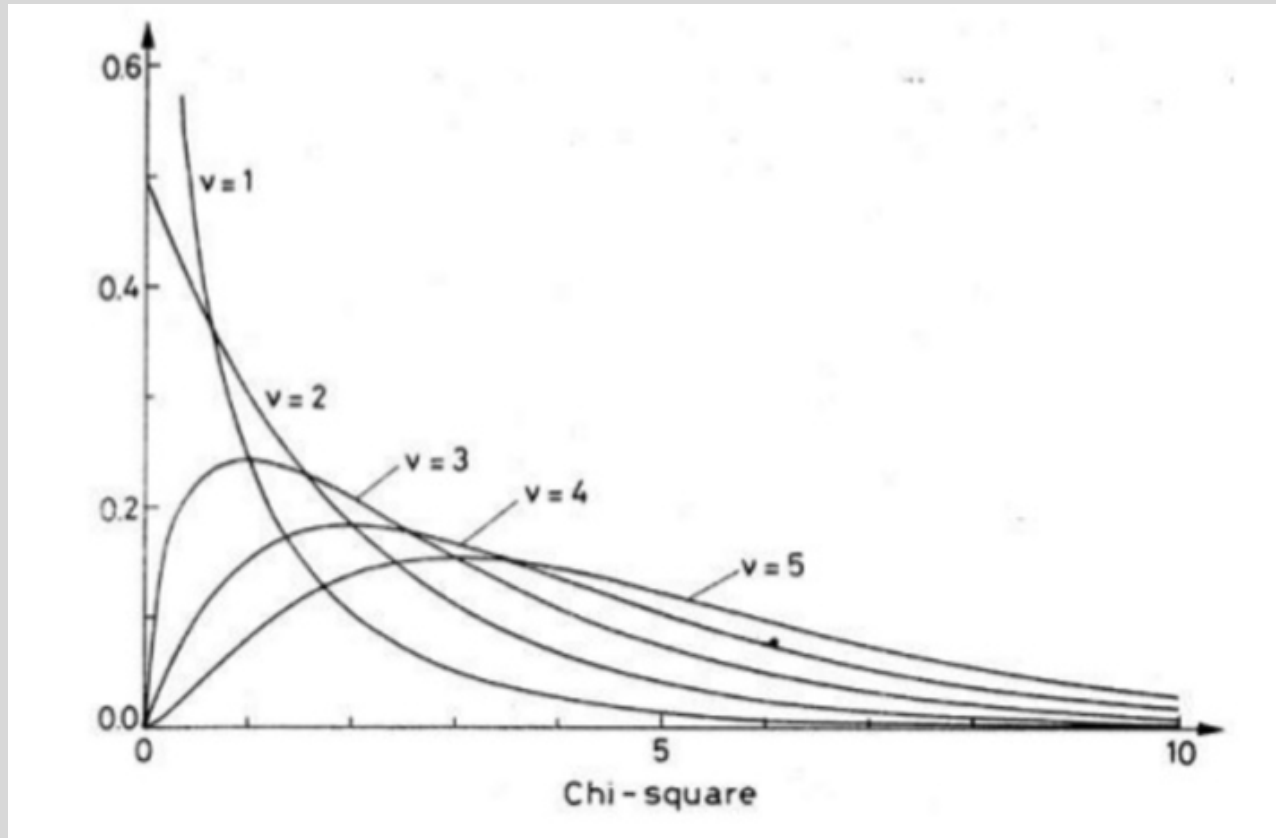
Chi²

- Chi² used to investigate whether distributions of categorical variables differ from one another
 - Can come from experimental, questionnaire, or observational research
 - Participants are categorised in some way
- Numbers only represent the category of which the participant is a member
 - Number of people in each category are counted
 - Data, therefore, takes the form of frequencies

Chi²

- “Not based on any assumption or distribution of variables”...
- But - follows a specific distribution known as *Chi² distribution*

Chi² distribution



- χ^2 distribution is not symmetrical and all values are positive. For each DF, there are asymmetric curves

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

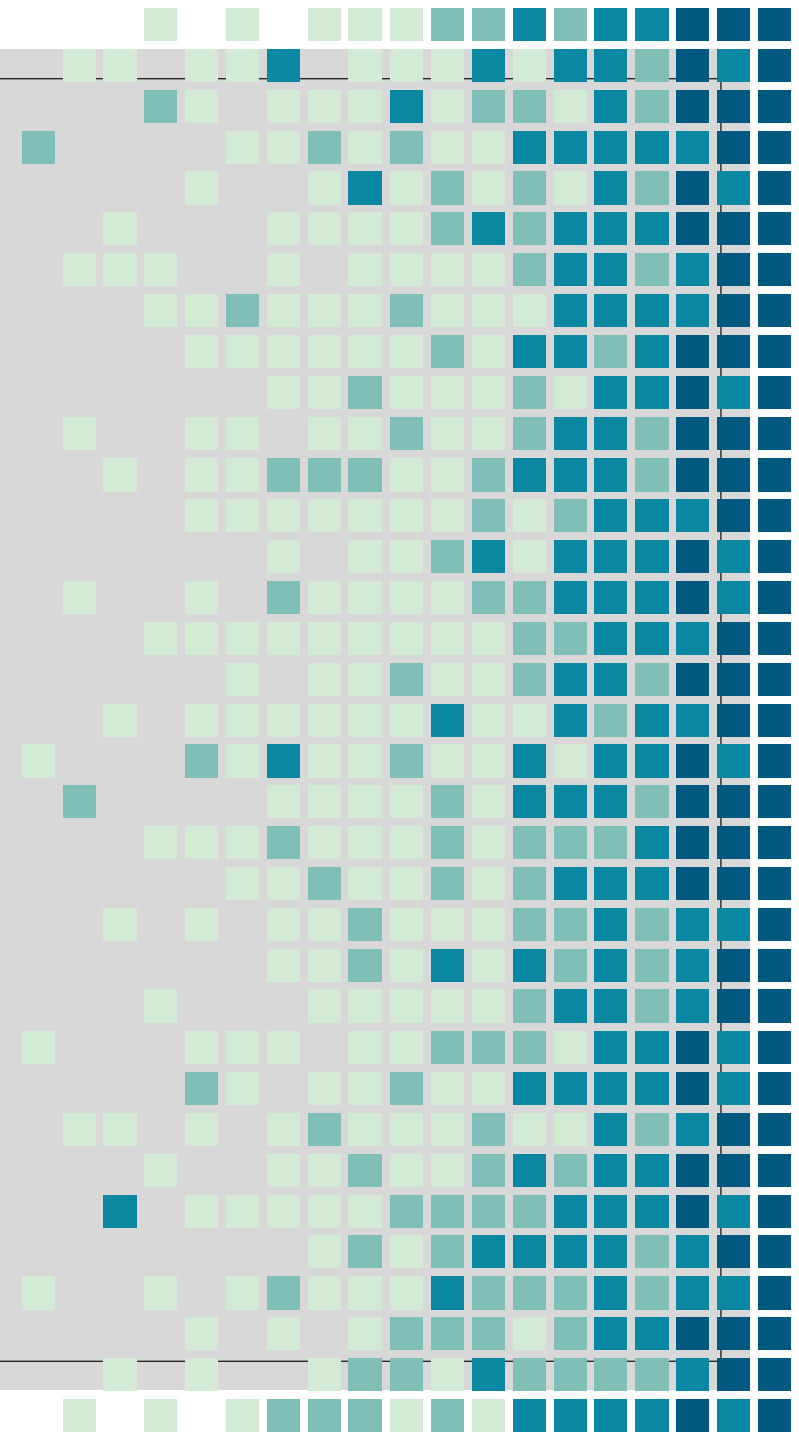
Chi²

- A large discrepancy = large X^2 value
- This indicates that the data do not fit the null hypothesis and the null hypothesis should be rejected
 - Experimental hypothesis therefore gets accepted

Chi²: Assumptions

- Data must be from a random sample from the population
- There is an expectation that no cells will have an expected frequency of less than 5
 - At the absolute minimum, up to 20% of cells can have an EF of less than 5; anymore and the analysis should not be used
 - Unless...
 - Some categories can be combined (if meaningful!)
 - More data is collected

Chi² for independence

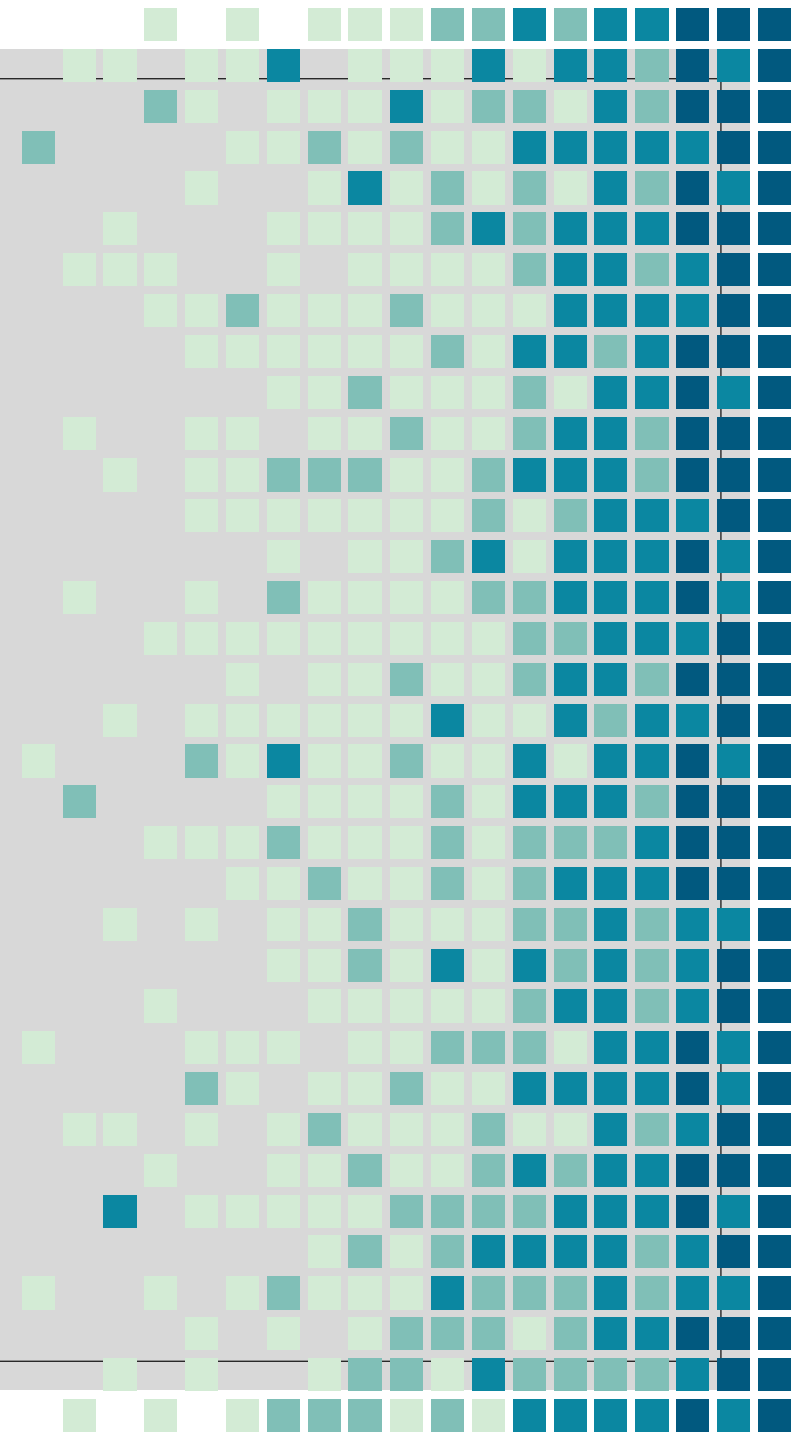


Chi² for independence

- Enables us to explain whether or not two attributes are associated
- If two distributions (observed and theoretical) are exactly alike, then $X^2 = 0$
 - But generally, due to sampling errors, X^2 is not equal to 0

Study example: Researcher
interested in voting
preferences from people in
different parts of Wales

Sent to random sample of 100 voters



Calculating Chi²

1. Make **contingency table**
2. Note observed frequencies (O) 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

1. Make contingency tables

	Remain	Leave	<i>Row frequency</i>
Rural			
Urban			
<i>Column frequency</i>			

2. Note observed frequencies (O) per cell

	Remain	Leave	<i>Row frequency</i>
Rural	45	15	
Urban	10	30	
<i>Column frequency</i>			N = 100

3. Note row and column frequencies

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

4. Determine expected frequencies (F_e)

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

	Remain	Leave	<i>Row frequency</i>
Rural	45	15	60
Urban	10	30	40
<i>Column frequency</i>	55	45	$N = 100$

4. Determine expected frequencies

	Remain	Leave	<i>Row frequency</i>
Rural	45 $F_e = 33$	15 $F_e = 27$	60
Urban	10 $F_e = 22$	30 $F_e = 18$	40
<i>Column frequency</i>	55	45	$N = 100$

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

5. Calculate χ^2 statistic

- Are the observed counts (our data) so different from the expected counts (that we have now calculated) that we can conclude a relationship between the two variables?
- We use the **χ^2 statistic** to compare each cell's observed counts to its respective expected count

SYMBOLS

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

X^2 = chi-square

Σ (Greek letter sigma) = the sum of

O = observed values

E = expected values

Superscript 2 (²) = square the attached value

5. Calculate χ^2 statistic

Stages of calculation:

Tip – do not attempt to do ALL the formula in one go!

Stage 1. Calculate any values in brackets

Stage 2. Square the calculated value (as per the superscript 2)

Stage 3. Do the division

Stage 4. Add up (sum) **all** values (for this we have 4) to get a final χ^2 statistic

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

Observed	Expected
45	33
15	27
10	22
30	18

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

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

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

$$X^2 = 4.364 + 5.333 + 6.545 + 8 = \mathbf{24.24}$$

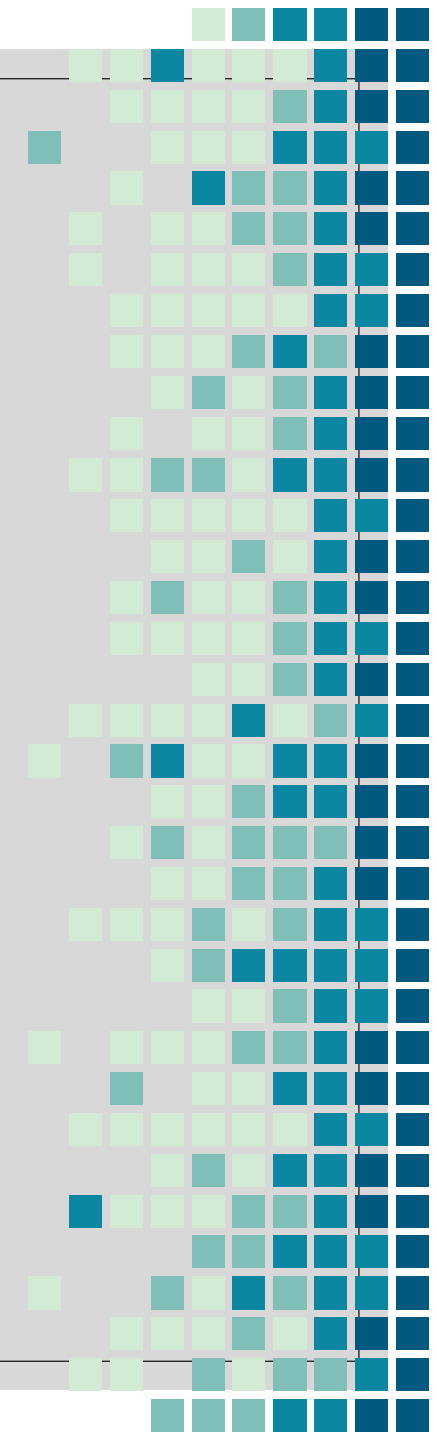
(reported to 2dp)

Stage 1. Brackets

Stage 2. Square

Stage 3. Divide

Stage 4. Sum



6. Work out DF

- DF refers to ‘degrees of freedom’; the number of data points that are free to vary given that overall summary values are known and remain constant
- DF is a number that determines the shape of our distribution, and for χ^2 is calculated as
- **DF =**
(number of rows – 1) x (number of columns – 1)

$$DF = (R-1) (C-1)$$

$$DF = (2-1) (2-1)$$

$$DF = 1 \times 1$$

$$\mathbf{DF = 1}$$

7. Determine whether sig. from critical values table

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

INTRODUCTION TO POPULATION GENETICS, Table D.1
© 2013 Sinauer Associates, Inc.

$$\begin{aligned} \text{DF} &= 1 \\ \text{Chi}^2 &= 24.24 \\ \text{Sig. value} &= 0.05 \end{aligned}$$

- Use critical values table and calculated statistics to conclude whether you can reject, or must accept, your null hypothesis
 - If value on critical value table is smaller than the value we have calculated, we can conclude there is a significant association and can reject the null hypothesis (or accept the research hypothesis)

**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

INTRODUCTION TO POPULATION GENETICS, Table D.1
© 2013 Sinauer Associates, Inc.

7. Determine whether sig. from critical values table

Our value of 24.24 is larger than the value of 3.84 shown here, so we can say that there is a significant association between where people live and their Brexit vote; people have different views on Brexit and these views are linked to where they live.

Effect Size

- The real/true effect of an association (if one has been found)
- For a chi-square analysis, the effect size can be calculated using Cramer's phi Φ , where X^2 is the final chi-square value calculated, and N is the total number of participants in the study.

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

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

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

$$\Phi = 0.49$$

Cohen (1988)
produced
effect size
conventions
for phi – see
textbook
which takes
into account df

Writing results for chi-square

- 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, $X^2 = (1, N = 100) = 24.24, p < 0.05$. 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.