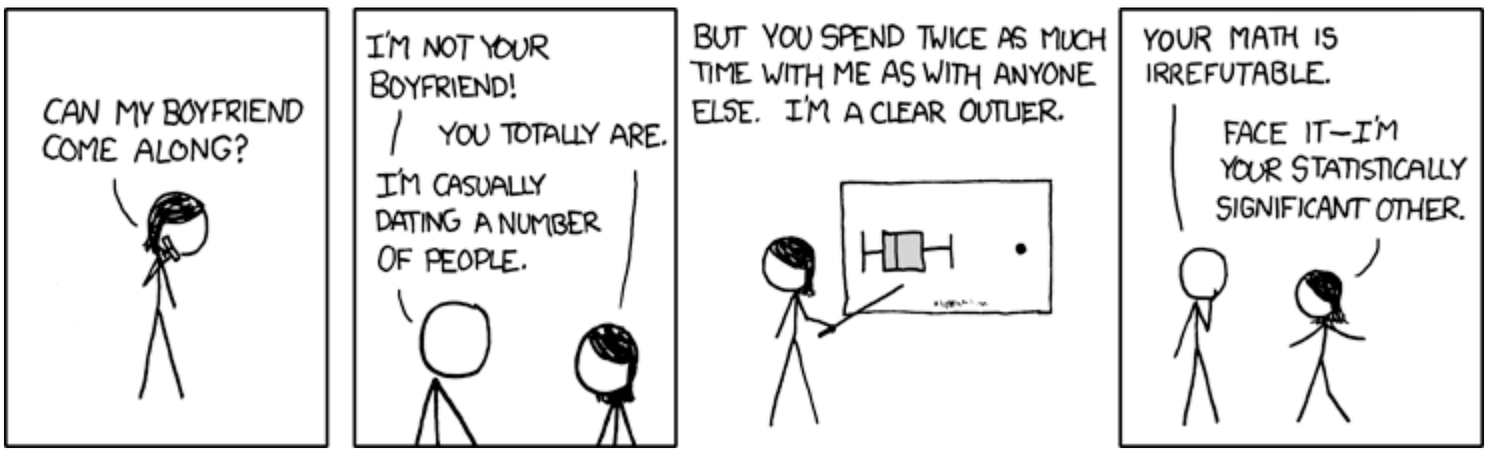
Research Methods for Political Science

Bivariate statistics: cross tables and chi-square



TRINITY COLLEGE DUBLIN
COLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH
THE UNIVERSITY OF DUBLIN

Dr. Thomas Chadefaux
Assistant Professor in Political Science
Thomas.chadefaux@tcd.ie

Bivariate statistics

- Bivariate: relationship between two variables
- Today: relationship between two nominal or ordinal variables
  - Cross tables
  - Chi-square

Cross table

Table: Turnout and home ownership

	Owner without a mortgage or loan	Owner with a mortgage or loan	Local authority tenant	Private tenant	Other
I voted in the election	91.1%	89.5%	80.0%	69.2%	81.2%
I did not vote in election	6.1%	7.3%	10.0%	22.0%	17.2%
I thought about voting but didn't	0.0%	1.0%	4.3%	3.3%	1.6%
I usually vote but didn't	2.9%	2.2%	5.7%	5.5%	0.0%

Rules for a crosstable

- Keep ordering for ordinal variables
- Independent variables in the columns
- Dependent variables in the rows
- Calculate **column** percentages
- Compare percentages across the **rows**

Table: Turnout and home ownership

	Owner without a mortgage or loan	Owner with a mortgage or loan	Local authority tenant	Private tenant	Other
I voted in the election	91.1%	89.5%	80.0%	69.2%	81.2%
I did not vote in election	6.1%	7.3%	10.0%	22.0%	17.2%
I thought about voting but didn't	0.0%	1.0%	4.3%	3.3%	1.6%
I usually vote but didn't	2.9%	2.2%	5.7%	5.5%	0.0%

%

Measures of association

- Generally these measure the strength of the relationship between two variables

- Which one to use depends on the measurement level
  - Categorical or ordinal -> chi-square based
  - Continuous (interval-ratio) -> correlation

Home ownership and voting

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	938	119	1057
		%	90.5%	73.9%	88.2%
	Did not vote	Count	99	42	141
		%	9.5%	26.1%	11.8%
Total		Count	1037	161	1198
		%	100.0%	100.0%	100.0%

Chi squared

Difference in the sample, can we generalize this to the population?

Chi squared

- **Observed** frequencies (f_o)
- **Expected** frequencies (f_e), if variables would not be related

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Observed frequencies

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	938	119	1057
	Did not vote	Count	99	42	141
Total		Count	1037	161	1198
Total (%)			0.87%	0.13%	100%

How likely is it that we obtained these numbers by chance?

That is: if there was no relationship between ownership and voting in the population, how likely is it we would get numbers which are so far away from what we would expect (or more extreme)?

Expected frequencies

			Owner or tenant		Total	Total (%)
			Owner	Tenant		
Vote in 2007 election	Did vote	Count	A	C	1057	88%
	Did not vote	Count	B	D	141	12%
Total		Count	1037	161	1198	100%
Total (%)			0.87%	0.13%	100%	

- If ownership and vote were not related, how many respondents should we expect in cell A?

Expected frequencies

			Owner or tenant		Total	Total (%)
			Owner	Tenant		
Vote in 2007 election	Did vote	Count	A = 88%*0.87%	C = 88%*0.13%	1057	88%
	Did not vote	Count	B = 12%*0.87%	D = 12%*0.13%	141	12%
Total		Count	1037	161	1198	100%
Total (%)			0.87%	0.13%	100%	

- If ownership and vote were not related, how many respondents should we expect in cell A?
  - If among all voters, 88% did vote, we would expect that among owners, also 88% would vote.
  - If among all owners, 87% did vote, we would expect that among voters, also 87% would vote.

Expected frequencies

			Owner or tenant		Total	Total (%)
			Owner	Tenant		
Vote in 2007 election	Did vote	Count	A	C	1057	88%
	Did not vote	Count	B	D	141	12%
Total		Count	1037	161	1198	100%
Total (%)			0.87%	0.13%	100%	

- If ownership and vote were not related, how many respondents should we expect in cell A?
- Expected frequency (f_e)= row margin$*\frac{\text{column margin}}{\text{total}}$
- $f_e = 1037 * \frac{1057}{1198}$
- f_e = 1037 * 0.88 = 914.9

Expected frequencies

			Owner or tenant		**Total**
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	A	C	1057
	Did not vote	Count	B	D	141
Total		Count	1037	161	1198

$$\text{Expected frequency } (f_e) = \frac{\text{row margin} * \text{colum margin}}{\text{total}}$$

Expected frequencies

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	A	C	1057
	Did not vote	Count	B	D	141
Total		Count	1037	161	1198

- B: f_e = 1037 * 141 / 1198 = 122.1
- C: f_e = 1057* 161 / 1198 = 142.1
- D: f_e = 141 * 161 / 1198 = 18.9

Expected frequencies

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	914.9	142.1	1057
	Did not vote	Count	122.1	18.9	141
Total		Count	1037	161	1198

Observed frequencies

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	938	119	1057
	Did not vote	Count	99	42	141
Total		Count	1037	161	1198

Expected frequencies

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	914.9	142.1	1057
	Did not vote	Count	122.1	18.9	141
Total		Count	1037	161	1198

Observed frequencies

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	938	119	1057
	Did not vote	Count	99	42	141
Total		Count	1037	161	1198

Expected frequencies

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	914.9	142.1	1057
	Did not vote	Count	122.1	18.9	141
Total		Count	1037	161	1198

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Cell A: $\frac{(f_o - f_e)^2}{f_e} = \frac{(938 - 914.9)^2}{914.9} = \frac{533.61}{914.9} = 0.58$

Observed frequencies

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	938	119	1057
	Did not vote	Count	99	42	141
Total		Count	1037	161	1198

Expected frequencies

			Owner or tenant		Total
			Owner	Tenant	
Vote in 2007 election	Did vote	Count	914.9	142.1	1057
	Did not vote	Count	122.1	18.9	141
Total		Count	1037	161	1198

Cell B: $\frac{(f_o - f_e)^2}{f_e} = \frac{(99-122.1)^2}{122.1} = \frac{533.61}{122.1} = 4.37$

Cell C: $\frac{(f_o - f_e)^2}{f_e} = \frac{(119-142.1)^2}{142.1} = \frac{533.61}{142.1} = 3.76$

Cell D: $\frac{(f_o - f_e)^2}{f_e} = \frac{(42-18.9)^2}{18.9} = \frac{533.61}{18.9} = 28.23$

Chi squared

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = .58 + 4.37 + 3.76 + 28.23 = 36.94$$

Great… but what does that mean?

Chi squared

- We need to compare the chi squared we **obtained** with the **critical** value for chi squared.
- If $\chi^2_{obtained} > \chi^2_{critical}$ we can conclude that it is unlikely that the relationship we found is just due to sampling error.

The critical value

- First, we need to set a **confidence level**, normally 95%
- This corresponds to a ***p* value** of 0.05 (1 – 95/100).
- Second, we need to know the **degrees of freedom**: df = (c-1)(r-1)

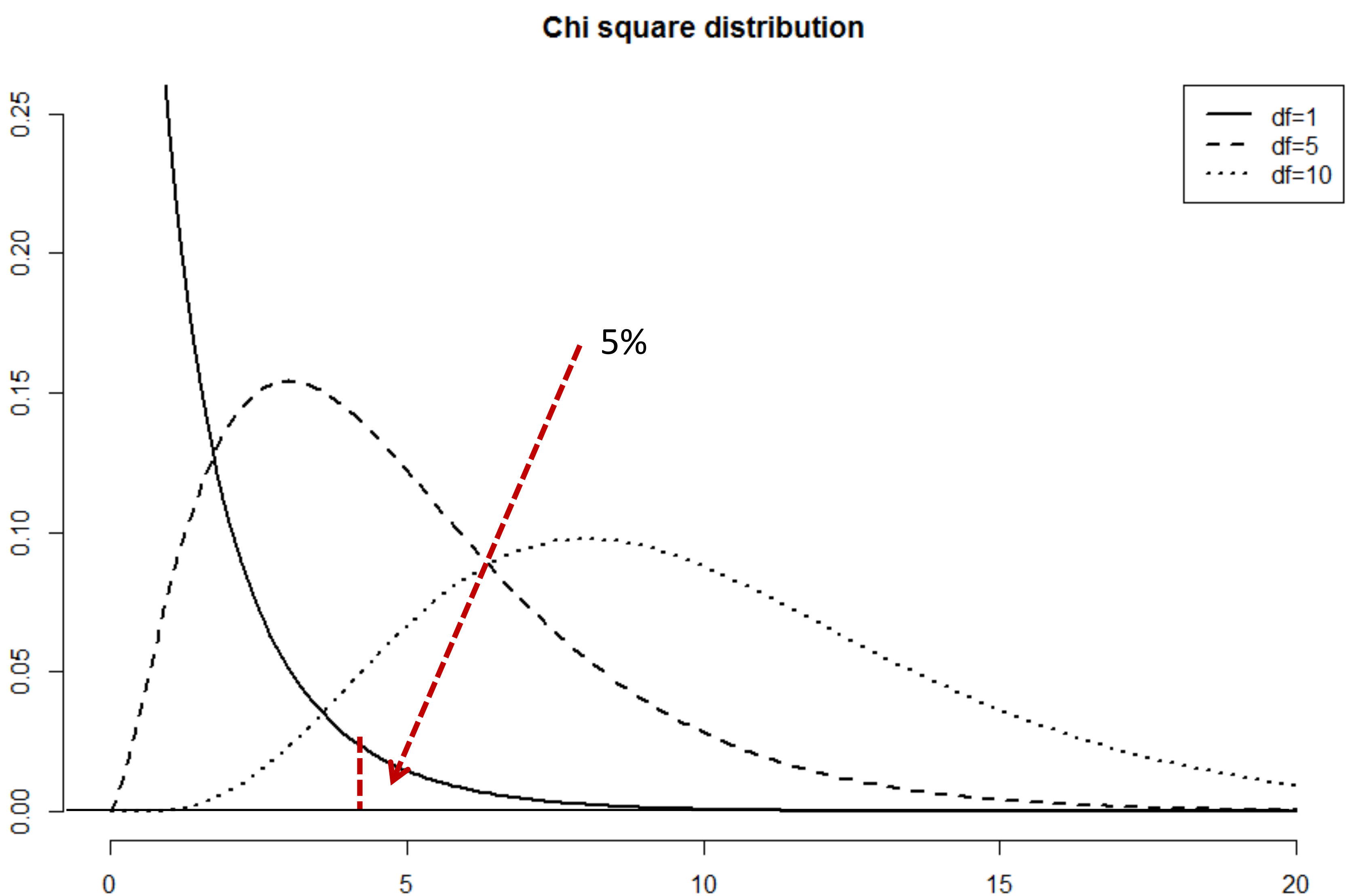
The critical value

In our example:

- The degrees of freedom:
  - 2 rows
  - 2 columns
  - df = (2 -1) * (2 – 1) = 1 * 1 = 1
- The critical value corresponding to df = 1 and p = 0.05 is found in Field, appendix A.4:

A.4. Critical values of the chi-square distribution

	p			
df	0.05	0.01	df	0.05
1	3.84	6.63	25	37.65
2	5.99	9.21	26	38.89
3	7.81	11.34	27	40.11
4	9.49	13.28	28	41.34
5	11.07	15.09	29	42.56
6	12.59	16.81	30	43.77
7	14.07	18.48	35	49.80


Chi square distribution
df=1
df=5
df=10
5%
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20

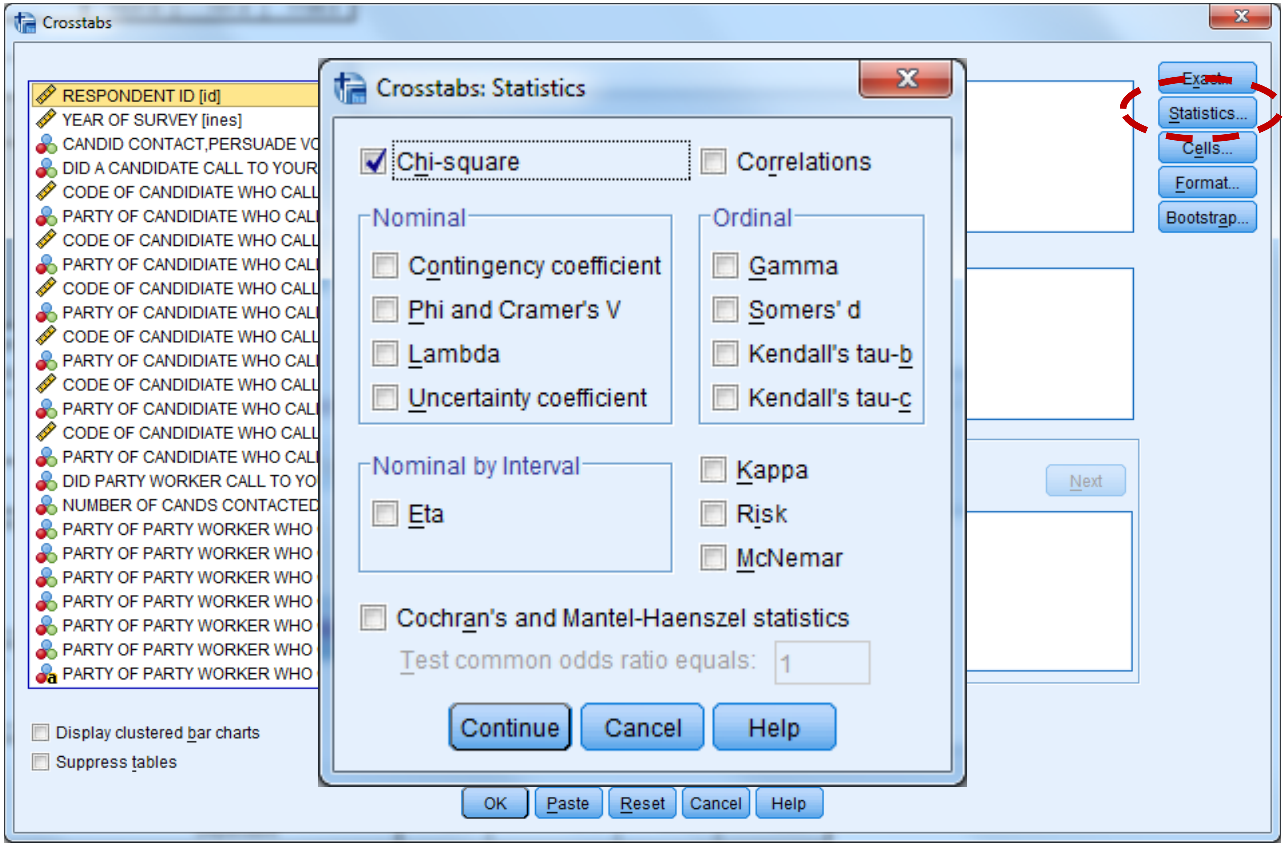
Comparing obtained and critical value

- $\chi^2_{obtained} = 36.94$
- $\chi^2_{critical} = 3.84$

- As $\chi^2_{obtained} > \chi^2_{critical}$ we conclude that there is a statistically significant relationship.

In SPSS

Analyze ... Descriptive Statistics ... Crosstabs



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Vote in 2007 election * Owner or tenant	1198	11.5%	9225	88.5%	10423	100.0%

Vote in 2007 election * Owner or tenant Crosstabulation

% within Owner or tenant

		Owner or tenant		Total
		Owner	Tenant	
Vote in 2007 election	Did vote	90.5%	73.9%	88.2%
	Did not vote	9.5%	26.1%	11.8%
Total		100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	36.715[a]	1	.000	
Continuity Correction[b]	35.140	1	.000	
Likelihood Ratio	29.949	1	.000	
Fisher's Exact Test				.000
Linear-by-Linear Association	36.685	1	.000	
N of Valid Cases	1198			

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is

b. Computed only for a 2x2 table

Chi squared

- If our N increases, our Chi-squared obtained will be larger. Thus: large N, more likely to find a statistically significant relationship
- If the number of categories increases, our degrees of freedom will increase, increasing Chi-squared critical. Thus: more categories, less likely to find a statistically significant relationship.

Assumptions of Chi squared

- Independent observations: each person, country, or other observation should only contribute to one cell in the cross table
- Expected frequencies should be greater than 5 in each cell. (Otherwise the sampling distribution of the Chi squared *statistic* does not follow a Chi squared *distribution*)

Strength of association

- Chi squared does not tell you *how strong* a relationship is, only whether it is statistically significant.
- If N is large, you are likely to find a significant relationship (but it might be a weak one).

- Solution: look at a measure of association, such as *Cramers' V.*

Cramer's V

- When your table is larger than 2x2, we should use Cramer's V (because Phi would never reach 0 in these cases):

- $V = \sqrt{\frac{\chi^2}{N*(\text{Minimum of } r-1, c-1)}}$

- The minimum of r – 1, c – 1, in our case is: the minimum of 2 – 1 and 2 – 1, which is 1.

Cramers' V

- If we find a Chi square of 80 for a 3 x 5 table, with N = 900.
- $V = \sqrt{\frac{\chi^2}{N*(\text{Minimum of } r-1, c-1)}}$
- $V = \sqrt{\frac{80}{900*(\text{Minimum of } 3-1, 5-1)}} = \sqrt{\frac{80}{900*2}}$
- $V = 0.21$

In SPSS

- Select Phi/Cramer's V in the 'Statistics' dialog when making a crosstable (Analyze ... Descriptive Statistics ... Crosstable).

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.175	.000
	Cramer's V	.175	.000
N of Valid Cases		1198	