Tutorial 09, Michaelmas Term

Research Methods for Political Science (PO3600)

Stefan Müller

12 December 2017

Trinity College Dublin
http://muellerstefan.net/research-methods

Session Outline

1. Chi Square Test
2. How accurate are elections polls?
3. Calculating confidence intervals for proportions

Chi Square Test

- A Chi Square Test determines if a sample data matches a population.
- Formula: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
- O: observed value; E: expected value
- Problem: calculation by hand takes long $\longleftrightarrow$ SPSS!

Chi Square Test: Example

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29				
Taurus	24				
Gemini	22				
Cancer	19				
Leo	21				
Virgo	18				
Libra	19				
Scorpio	20				
Sagittarius	23				
Capricorn	18				
Aquarius	20				
Pisces	23				

Figure: Source: http://www.statisticshowto.com/probability-and-statistics/chi-square/.

Chi Square Test: Example

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333			
Taurus	24	21.333			
Gemini	22	21.333			
Cancer	19	21.333			
Leo	21	21.333			
Virgo	18	21.333			
Libra	19	21.333			
Scorpio	20	21.333			
Sagittarius	23	21.333			
Capricorn	18	21.333			
Aquarius	20	21.333			
Pisces	23	21.333			

Chi Square Test: Example

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667		
Taurus	24	21.333	2.667		
Gemini	22	21.333	0.667		
Cancer	19	21.333	-2.333		
Leo	21	21.333	-0.333		
Virgo	18	21.333	-3.333		
Libra	19	21.333	-2.333		
Scorpio	20	21.333	-1.333		
Sagittarius	23	21.333	1.667		
Capricorn	18	21.333	-3.333		
Aquarius	20	21.333	-1.333		
Pisces	23	21.333	1.667		

Chi Square Test: Example

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	
Taurus	24	21.333	2.667	7.112889	
Gemini	22	21.333	0.667	0.44889	
Cancer	19	21.333	-2.333	5.442889	
Leo	21	21.333	-0.333	0.110889	
Virgo	18	21.333	-3.333	11.108889	
Libra	19	21.333	-2.333	5.442889	
Scorpio	20	21.333	-1.333	1.776889	
Sagittarius	23	21.333	1.667	2.778889	
Capricorn	18	21.333	-3.333	11.108889	
Aquarius	20	21.333	-1.333	1.776889	
Pisces	23	21.333	1.667	2.778889	

Chi Square Test: Example

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

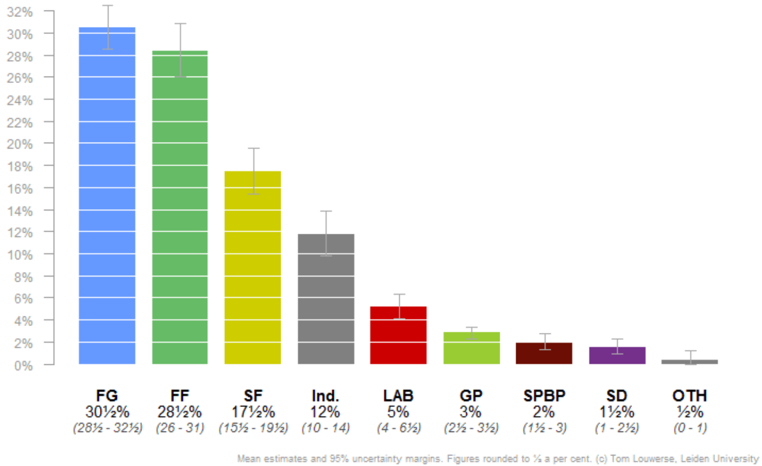
Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457
					5.094017203

Test Chi Square Hypothesis: Steps

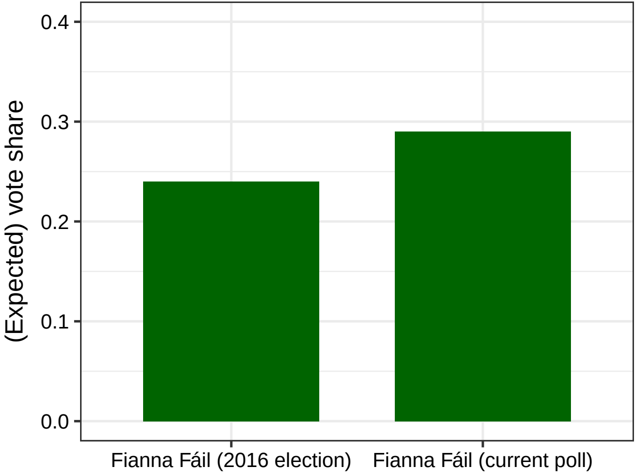
Question: Difference in sample, can we generalise this to the population?

1. 11 Degrees of Freedom; Chi square test statistic of 5.094
2. Take the chi-square statistic and find value in chi-square table
3. closest value for df=11 and 5.094 is between .900 and .950.
4. Decide whether this p-value supports or rejects the null hypothesis. This very large p-value means that the null hypothesis cannot be rejected.

Irish Polling Indicator



Confidence intervals for proportions



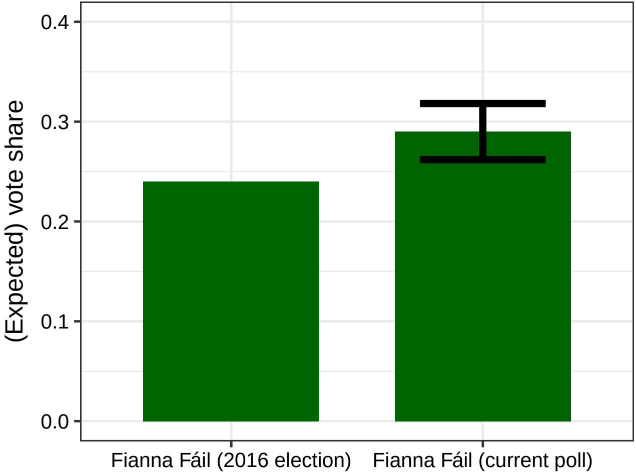
Confidence intervals for proportions

Is there a statistically significance in support for FF in December 2017 compared to the election in February 2016?

What we know:

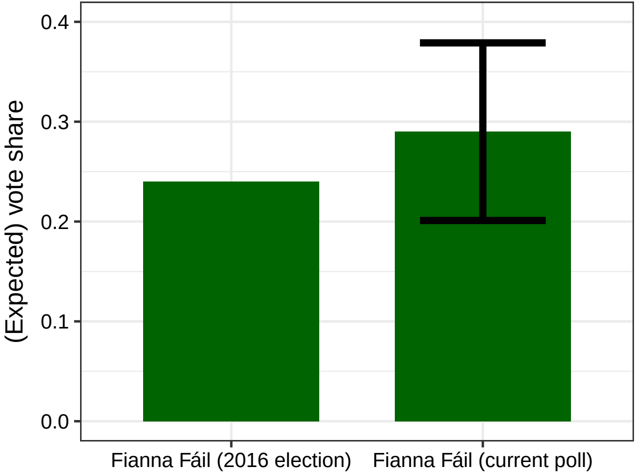
- 2016 FF election result: 24%; FF standing in latest poll (p) = 29%; number of respondents in poll = 1001; 95 percent confidence interval (z) = 1.96
- $CI = p \pm z * \sqrt{(p \times (1 - p)/N}$
- $CI_{low} = 0.29 + 1.96 * \sqrt{(0.29 \times (1 - 0.29)/1001} = 0.26$
- $CI_{high} = 0.29 - 1.96 * \sqrt{(0.29 \times (1 - 0.29)/1001} = 0.31$

Confidence intervals for proportions



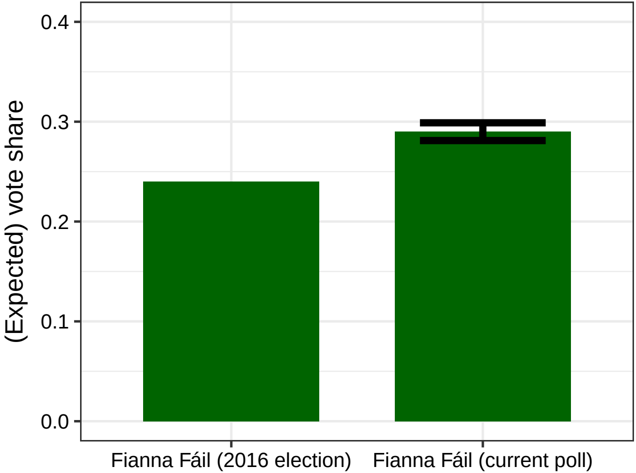
Confidence intervals for proportions

How does plot look like when we only ask 30 Irish voters?



Confidence intervals for proportions

. . . and for 10,000 respondents?




KEEP
CALM
AND
REPORT
CONFIDENCE INTERVALS

Research Proposal

Team up and discuss (some of) the following aspects:

1. Research question + relevance
2. Theoretical argument + hypothesis
3. Type of data + operationalisation of variables
4. Ways of analysing your data

Research Proposal

Discuss with your neighbour:

1. What is the puzzle you address?
2. Why is the question relevant?
3. What are your key dependent and independent variables?
4. What data will you use?

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 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

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

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

Interesting, 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 cricical 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 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
1	$\boxed{3.84}$	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48

df	0.05
25	37.65
26	38.89
27	40.11
28	41.34
29	42.56
30	43.77
35	49.80

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.