SO5041 Unit 11:

Brendan Halpin, Sociology

Autumn 2023/4

SO5041 Unit 11

Two new distributions: t and χ^2

Today

- The χ^2 test for association in tables (chi-square)
- The t-distribution: replaces the normal for small samples

SO5041 Unit 11

Samples and tables

Association in tables

- When we make a table, we read it for association by looking at percentages
- Either the row percentages, up and down the columns
- Or the column percentages across the rows
- (If one variable happens before or causes the other, percentages within that variable are easier to interpret)

ESS: selected countries and attitude, row %

```
. tab cntry freehms, row nokey
```

Country	Gays and lesbians free to live life as they wish: Agree str	Agree	Neither a	Disagree	Disagree	Total
DE	1,135	926	144	78	56	2,339
	48.53	39.59	6.16	3.33	2.39	100.00
FR	1,329	450	119	47	49	1,994
	66.65	22.57	5.97	2.36	2.46	100.00
GB	1,057	884	176	49	28	2,194
	48.18	40.29	8.02	2.23	1.28	100.00
IE	1,001	970	126	53	34	2,184
	45.83	44.41	5.77	2.43	1.56	100.00
Total	4,522	3,230	565	227	167	8,711
	51.91	37.08	6.49	2.61	1.92	100.00

Source: European Social Survey, Wave 9 (2018/9)

sociology
UNIVERSITY OF LIMERICK

ESS: selected countries and attitude, column %

```
. tab cntry freehms, col nokey
```

Country	Gays and lesbians free to live life as they wish: Agree str	Agree	Neither a	Disagree	Disagree	Total
DE	1,135	926	144	78	56	2,339
	25.10	28.67	25.49	34.36	33.53	26.85
FR	1,329	450	119	47	49	1,994
	29.39	13.93	21.06	20.70	29.34	22.89
GB	1,057	884	176	49	28	2,194
	23.37	27.37	31.15	21.59	16.77	25.19
IE	1,001	970	126	53	34	2,184
	22.14	30.03	22.30	23.35	20.36	25.07
Total	4,522	3,230	565	227	167	8,711
	100.00	100.00	100.00	100.00	100.00	100.00

Detecting association

- If these percentages differ from each other (or from the row/col total percentages) we say there is association
- The distribution of one variable depends on the other value of the other
- But how big do differences need to be before we say it's a meaningful level of association?
- NB: A sample from a population with no association will have small percentage differences just by chance!

Fake data with no association, row %

```
. tab c1 freehms, row chi nokey
```

c1	Gays and lesbians free to live life as they wish: Agree str	Agree	Neither a	Disagree	Disagree	Total
DE	1,206	889	147	57	40	2,339
	51.56	38.01	6.28	2.44	1.71	100.00
FR	992	755	145	55	47	1,994
	49.75	37.86	7.27	2.76	2.36	100.00
GB	1,146	797	146	63	42	2,194
	52.23	36.33	6.65	2.87	1.91	100.00
IE	1,178	789	127	52	38	2,184
	53.94	36.13	5.82	2.38	1.74	100.00
Total	4,522	3,230	565	227	167	8,711
	51.91	37.08	6.49	2.61	1.92	100.00

Pearson chi2(12) = 13.3692 Pr = 0.343

sociology
UNIVERSITY OF LIMERICK

Problem: how small is small

- A table that logically cannot have a link between nationality and attitude has small percentage differences just by chance
- We need a rule to decide
  - how small is small enough to say there is no association?
  - how big is big enough to count as evidence of association?
- Answer: the χ^2 test (chi-square)

SO5041 Unit 11

The χ^2 test

The χ^2 test for independence in tables

- The χ^2 test for association in tables
- Independence: no association between two variables
  - pattern of row percentages the same in all rows
  - pattern of column percentages the same in all columns
- Even if independence holds in the population, sampling variability leads to differences in percentages
- How big can the differences be before we can be convinced that there is really association in the population?

Compare “observed” with “expected”

- Method: compare the real table (“observed”) with hypothetical table under independence (“expected”)
- Summarise the difference into a single figure (χ^2 statistic, chi-sq)
- Compare χ^2 statistic with known distribution
- . . . What is the probability of getting a sample statistic “at least this big” by simple sampling variability *if independence holds in the population*?

"Expected" ⇒"independence"

- The "expected" table has the same row and column totals, but the cell values are such that the percentages are the same as in the total row and column:

$$n_{ij} = \frac{R_i C_j}{T}$$

- For each cell we summarise the difference between observed (O) and expected (E) values as

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

- The summary for the table as a whole is the sum of this quantity across all cells:

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

Illustration

Demonstration with spreadsheet

X^2 has a χ^2 distribution

- This statistic has a known distribution, the χ^2 distribution
- That is, if we take a large number of samples from a population where there is no association, and calculate the statistic, they will have a distribution in a known form, and we can calculate the probability of finding a value “at least as large as” any given number
- Depends only on the “degrees of freedom”: number of rows minus one times the number of columns minus one:

$$df = (r - 1)(c - 1)$$

χ^2 distribution

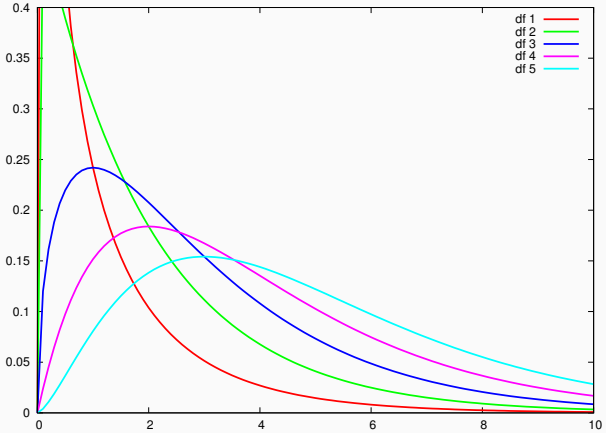


Figure 1: The χ^2 Distribution for various degrees of freedom

App

This app draws the χ^2 distribution for different degrees of freedom, and calculates the proportion of the distribution above a given χ^2 statistic.

http://teaching.sociology.ul.ie:3838/apps/chidist

With Stata

```
. tab cntry freehms, row nokey chi
```

Country	Gays and lesbians free to live life as they wish: Agree str	Agree	Neither a	Disagree	Disagree	Total
DE	1,135	926	144	78	56	2,339
	48.53	39.59	6.16	3.33	2.39	100.00
FR	1,329	450	119	47	49	1,994
	66.65	22.57	5.97	2.36	2.46	100.00
GB	1,057	884	176	49	28	2,194
	48.18	40.29	8.02	2.23	1.28	100.00
IE	1,001	970	126	53	34	2,184
	45.83	44.41	5.77	2.43	1.56	100.00
Total	4,522	3,230	565	227	167	8,711
	51.91	37.08	6.49	2.61	1.92	100.00

```
Pearson chi2(12) = 294.6550   Pr = 0.000
```

SO5041 Unit 11

The t-distribution

The t-distribution

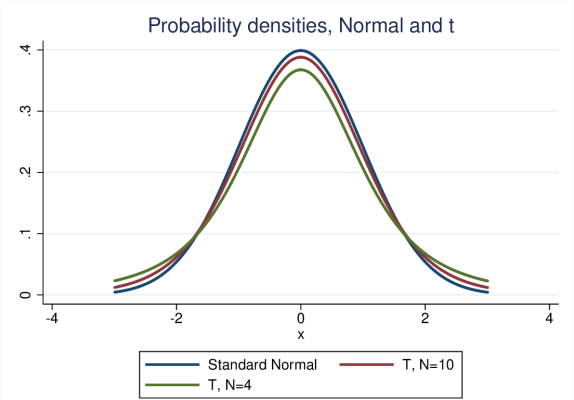
- We have used the Standard Normal Distribution to make confidence intervals around sample means and sample proportions
- The depends on the Central Limit Theorem:

  A sufficiently large sample drawn from any population will have a normal distribution, even when the population distribution is not normal

- In practice, for small samples the approximation to the normal distribution is not good, so we use a related distribution called the t-distribution

Student's *t*

- The t-distribution is like the normal distribution but wider
- The bigger the sample the closer it is to the normal distribution
- Small samples give much fatter tails
- It makes the CI a little wider for smaller sample sizes



Student's identity

- It is known as "Student's t", after the Guinness statistician who invented it
- William Sealey Gosset (1876–1937)
- Guinness were wary of losing trade secrets so publication of research was not allowed
- Gosset published anonymously, as "Student"

Sample size matters: degrees of freedom

- If the sample is small and we are calculating, say, a mean, the normal distribution underestimates the uncertainty
- That is, repeated samples from the same population will vary more widely
- The true sampling distribution of the mean is wider
  - depends on degrees of freedom which is N-1

Online web applets

- See web applets:
  - Standard Normal Distribution: http://teaching.sociology.ul.ie:3838/apps/snd
  - t-Distribution: http://teaching.sociology.ul.ie:3838/apps/tdist
- More spread at small N, more like normal distribution at high N
- Use the app to see P above/below a t value, or inside/outside ± t

Worked example

- Let's use the same example as in lecture 12, transport spending:
  - sample mean: €34.658
  - sample standard deviation: €10.123
- But this time, let's say the sample was 20
- This makes the standard error bigger:

$$\frac{10.123}{\sqrt{20}} = 2.264$$

Calculating the interval

- CI: $\bar{X} \pm t_{0.95,19} \times SE$
- We can use the web-app to find the t-value for 95% confidence
- DF = 20 - 1 = 19

Calculating the interval

- CI: $\bar{X} \pm t_{0.95,19} \times SE$
- We can use the web-app to find the t-value for 95% confidence
- DF = 20 - 1 = 19

- t = 2.093
- Interval:

$$34.658 \pm 2.093 \times 2.264$$

$$29.920 \leftarrow 34.658 \rightarrow 39.396$$

A wider interval

- A much wider interval than in Lecture 12, but that is mostly because of the small sample size making the standard error much bigger
- However, it is also a wider interval than if we used the SND
  - Correct:

$$29.920 \leftarrow 34.658 \rightarrow 39.396$$

  - Using 1.96 instead of 2.093 (incorrect)

$$30.221 \leftarrow 34.658 \rightarrow 39.095$$

Why it matters

- The interpretation of a 95% confidence interval is:
  - For 95% of samples, the true value will fall in the interval
  - We say we are 95% confident the true value falls in the interval
- If we use z = 1.96, our interval is too narrow and the true value will fall outside the interval more than 5% of the time

When to use the t-Distribution

- If working by hand (e.g., doing an exam question):
  - If N is above about 60-100, use the Normal Distribution
  - Otherwise use the t-distribution
- If using a computer, always use t
  - If N is small, it is more correct
  - If N is large, it makes no difference
- Stats programs (Stata, R, SPSS, etc) will always use t
- Note t is relevant only for quantities like means
  - Don’t use for proportions

SO5041 Unit 11

Spreadsheet exercise

Spreadsheet exercise

- We now use a spreadsheet to generate a confidence interval step by step
- We will calculate the mean and the standard deviation for 16 observations
- Then calculate the standard error and create a confidence interval

Data

16 observations have been collected as follows:

36	54	25	54
84	44	98	19
57	27	97	51
60	13	68	81

Formulas

- Mean: $\bar{X} = \frac{\sum X_i}{N}$
- Standard deviation: $s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N-1}}$
- Standard error: $se = \frac{s}{\sqrt{N}}$
- CI: $\bar{X} \pm t \times se$