SO5041 Unit 11:

Brendan Halpin, Sociology

Autumn 2019/0

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" $\Rightarrow$ "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:

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

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:

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

Reading the χ^2 distribution table

- To read the table, we go to the row corresponding to the degrees of freedom, and read across until we get to the column with our chosen probability level (say 0.05)
- If our χ^2 is bigger than the table value, then there is at most one chance in 20 (i.e., 0.05) that it has arisen by sampling variability, if the population has no association

χ^2 distribution

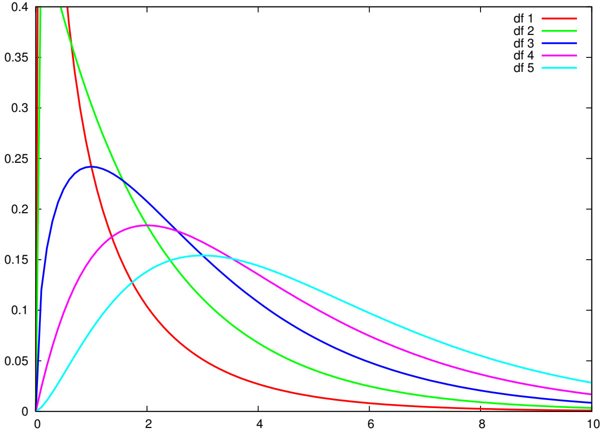

Figure: The χ^2 Distribution for various degrees of freedom

Table of the χ^2 distribution (chi-sq)

Values of the χ^2 statistic for various degrees of freedom and areas under the right tail. (see full table)

Degrees of Freedom	Area under right tail				
	0.100	0.050	0.025	0.010	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188

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 and flatter
- The bigger the sample the closer it is to the normal distribution
- 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 was banned

Degrees of freedom

- In practice, we first determine the “degrees of freedom”: this is the sample size minus 1
- Then we choose our probability level – a 95% CI has a probability level of 5% or 0.05
- Then we read off the t-value: this is used in constructing CIs exactly as the z-value from the Standard Normal Distribution

Confidence interval with t

- If the sample is large, Central Limit Theroem says sampling distribution is normal: use $\bar{X} \pm z \times SE$ for CI
- If the sample is small, sampling distribution is wider than normal: use $\bar{X} \pm t \times SE$

Table of the t-distribution (part)

Table of the Student's t Distribution

Two-tailed probability

Degrees of Freedom	Probability level (Area under both tails)				
	0.10	0.05	0.025	0.01	0.005
1	6.314	12.706	25.452	63.657	127.321
2	2.920	4.303	6.205	9.925	14.089
3	2.353	3.182	4.177	5.841	7.453
4	2.132	2.776	3.495	4.604	5.598
5	2.015	2.571	3.163	4.032	4.773
6	1.943	2.447	2.969	3.707	4.317
7	1.895	2.365	2.841	3.499	4.029
8	1.860	2.306	2.752	3.355	3.833
9	1.833	2.262	2.685	3.250	3.690
10	1.812	2.228	2.634	3.169	3.581
. . .					

(see full table)

A worked example

- Example: $X = (9, 2, 5, 5, 13, 13, 5, 5, 1, 4)$

A worked example

- Example: $X = (9, 2, 5, 5, 13, 13, 5, 5, 1, 4)$
- $\bar{X} = 6.2$

A worked example

- Example: $X = (9, 2, 5, 5, 13, 13, 5, 5, 1, 4)$
- $\bar{X} = 6.2$
- $s = 4.158$

A worked example

- Example: $X = (9, 2, 5, 5, 13, 13, 5, 5, 1, 4)$
- $\bar{X} = 6.2$
- $s = 4.158$
- $SE = 1.315$

A worked example

- Example: $X = (9, 2, 5, 5, 13, 13, 5, 5, 1, 4)$
- $\bar{X} = 6.2$
- $s = 4.158$
- $SE = 1.315$
- 95% confidence, 9 degrees of freedom: $t_{0.05,9} = 2.262$

A worked example

- Example: $X = (9, 2, 5, 5, 13, 13, 5, 5, 1, 4)$
- $\bar{X} = 6.2$
- $s = 4.158$
- $SE = 1.315$
- 95% confidence, 9 degrees of freedom: $t_{0.05,9} = 2.262$
- Low: $6.2 - 2.262 \times 1.315 = 3.226$

A worked example

- Example: $X = (9, 2, 5, 5, 13, 13, 5, 5, 1, 4)$
- $\bar{X} = 6.2$
- $s = 4.158$
- $SE = 1.315$
- 95% confidence, 9 degrees of freedom: $t_{0.05,9} = 2.262$
- Low: $6.2 - 2.262 \times 1.315 = 3.226$
- High: $6.2 + 2.262 \times 1.315 = 9.174$