

The Chi Square Test and Measures of Association

EDP 613

Week 11

A Note About The Slides

Currently the equations do not show up properly in Firefox. Other browsers such as Chrome and Safari do work.

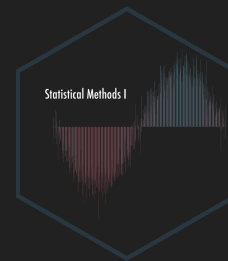


Independence

Two variables that have no association with each other are **statistically independent**.



Frequencies



- **expected frequencies**

written f_e

what you would *expect* in a bivariate table if two variables were statistically independent

only assumption: the null hypothesis is true

calculated by

$$f_e = \frac{\text{column marginal} \cdot \text{row marginal}}{\text{total sample size}}$$

- **observed frequencies**

written f_o

what you would *observe* in a bivariate table given what you have

calculated by you or given

Chi-Square Test



written χ^2 .

assumes *random sampling*

Is an inferential test to find significant relationships between two variables.

Calculated by

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

with

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

Example: Social Media

The percent of people using at least one social media outlet is given below by age groups

In 2011:

Age	Portion
18 - 29	820
30 - 49	590
50 - 64	360
65+	120

In 2021:

Age	Responses
18 - 29	840
30 - 49	810
50 - 64	730
65+	450

- Test the assumption that *users are equally likely* to be in each of the four age groups listed.
- Which age group contributes the largest amount to the test statistic?

Source: *Pew Research Center: Social Media Fact Sheet*

Example: Solution for 2011



a. We have

H_0 : Users are equally likely to be in each of the five groups listed

H_1 : Users are NOT equally likely to be in each of the five groups listed

Step 1: Find N

We have $820 + 590 + 360 + 120 = 1890$ total responses

If the distribution was uniform across all four categories, we would expect that each had $1890/4 \approx 472$ respondents

Step 2: Calculate the χ^2 statistic



Age	Responses	χ^2
18 - 29	820	$\frac{(820 - 1890)^2}{1890} \approx 605.767$
30 - 49	590	$\frac{(590 - 1890)^2}{1890} \approx 894.180$
50 - 64	360	$\frac{(360 - 1890)^2}{1890} \approx 1238.571$
65+	120	$\frac{(120 - 1890)^2}{1890} \approx 1657.619$

with the total

$$605.767 + 894.180 + 1238.571 + 1657.619 = 4396.137$$

and

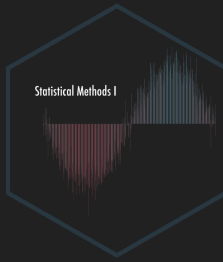
$$df = 4 - 1 = 3$$

Step 3: Make a Decision

In Appendix D

- Look at $df = 3$
- $\chi^2 = 4396.137 < \text{the greatest } p\text{-value so } p < 0.001$
- We reject H_0 implying that

respondents are not equally likely to be in each of the four age ranges listed



b .

- 65+ contributes the greatest amount to the sum for the test statistic
- The observed count is much smaller than expected



Example: Solution for 2021

We have

H_0 : Users are equally likely to be in each of the five groups listed

H_1 : Users are NOT equally likely to be in each of the five groups listed



Step 1: Find N

We have $840 + 810 + 730 + 450 = 2830$ total responses

If the distribution was uniform across all four categories, we would expect that each had $2830/4 \approx 707$ respondents



Step 2: Calculate the χ^2 statistic



Age	Responses	χ^2
18 - 29	840	$\frac{(840 - 2830)^2}{2830} \approx 1399.329$
30 - 49	810	$\frac{(810 - 2830)^2}{2830} \approx 1441.837$
50 - 64	730	$\frac{(730 - 2830)^2}{2830} \approx 1558.304$
65+	450	$\frac{(450 - 2830)^2}{2830} \approx 2001.555$

with the total

$$1399.329 + 1441.837 + 1558.304 + 2001.555 = 6401.025$$

and

$$df = 4 - 1 = 3$$

Step 3: Make a Decision

In Appendix D

- Look at $df = 3$
- $\chi^2 = 6401.025 < \text{the greatest } p\text{-value so } p < 0.001$
- We reject H_0 implying that

respondents are not equally likely to be in each of the four age ranges listed



That's it. Take a break before our R session!

