Contingency Table and Correlations

Bolun

12/1/2020

Warming up and Reviewing

- The basic logic of hypothesis test.
- What can you say based on a hypothesis test.
- Some reflection on the p-values.
- From a single variable to the relations between different variables.

The Idea of Correlation

Correlation is one of the most common relations between two variables. It is **any** statistical relationship between two random variables.

It is a statistical relationship. Thus, it is conceptually different from causality. Usually, social science research would like to step into the realm of causality. In quantitative research, we see a robust sub-field focusing on causal inference. But still, causality identification is more about research design.

Group Discussion

- People usually claim that correlation is different from causality. How do you understand it? Could you give an example that confuses correlation and causality?
- In your current knowledge, how would you address such a confusion?

Contingency table

- Contingency table is a tool to represent and examine relations between two categorical variables.
- In terms of format, it is no different from the table we studied previously in the previous sections.
- An example:

sex	demo	indep	repub	total
Females	573	516	422	1511
Males	386	475	399	1260
Total	959	991	821	2772

- Usually, the columns should be the dependent variable you suggest.

Conditional distributions

- We can in term calculate the row percentage, which is called conditional distribution.

sex	demo	indep	repub	total
Females	38%	34%	28%	100%
Males	31%	38%	32%	101%

- If the two variables are independent from each other, we should see similar row percentage in each columns.

But how to test it statistically?

- We guessed that there is a relationship between the gender and party identification.
- But to what extent we are sure about it?
- Statisticians found Chi-square statistics for us.

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

- O_i is the observed value and E_i is the expected value.
- review: how do you calculate the expected value of each cell?

Chi-square test

- The Chi-square statistics follows a chi-square distribution with a degree of freedom
  $(number\ of\ rows - 1)(number\ of\ columns - 1)$
- We won't simulate this, instead, we use functions to calculate the critical area.

Group Activity

- Find two categorical variables in GSS.
- Produce a contingency table/cross-tabulation.
- Calculate the Chi-square and test it with probable degree of freedom

Direction of the Correlation

- For 2*2 table, for example, two independent events, we can know the direction of the correlation using the odds ratio

Review

- Now we get to know whether there is a correlation between two variables.
- What we still do not know: the direction and the intensity of such correlation.
- OLS regression, and regression in general would help us to approach it.