Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 10. Association between qualitative variables. (07 February)
Alla Tambovtseva

Independent events and random variables: recall

- Events A and B are independent if the following equality holds:

$$P(A \cap B) = P(A) \times P(B).$$

- Random variables X and Y are independent if the following equality holds for every pair of values (x, y):

$$P(X = x \cap Y = y) = P(X = x) \times P(Y = y).$$

Consider the example suggested below.

Example 1. Here is the joint distribution of two random variables X and Y:

X\Y	0	1	2
-1	0.1	0.5	0.1
3	0	0.15	0.15

Let us write out marginal distributions of X and Y. So as to do it, we have to calculate corresponding probabilities, just to sum them by rows and columns. Look:

X	-1	3
p	0.7	0.3

Y	0	1	2
p	0.1	0.65	0.25

Now let us check the equality stated avove. First, we will take the following pair: $X = -1$ and $Y = 0$. From the joint distribution we can see that

$$P(X = -1 \cap Y = 0) = 0.1.$$

And how about probabilities from marginal distributions? Look:

$$P(X = -1) = 0.7 \text{ and } P(Y = 0) = 0.1.$$

Check the equality.

$$P(X = -1 \cap Y = 0) \neq P(X = -1) \times P(Y = 0),$$

so the condition for independence does not hold for at least one pair of values of X and Y. Hence, random variables X and Y are not independent.

This idea of independence will be helpful for our further discussion of a chi-squared test.

Contingency tables

What is a contingency table? It is a table of joint frequencies that describe the relationship between two nominal (categorical, qualitative) variables.

Example 2. A simple example of 2×2 contingency table that shows the relationship between people's propensity to vote for candidates A and B and their sex:

	Male	Female
Candidate A	40	25
Candidate B	10	55

So, 40 males and 25 females are going to vote for the candidate A, 10 males and 55 females are going to vote for the candidate B.

This table contains frequencies that are calculated from our empirical data, e.g. from the results of a real survey. These frequencies are usually called *observed frequencies*.

What should we do if we want to decide whether people's propensity to vote for a particular candidate depends on their sex? Probably, we should invent a special statistical measure that will help us to evaluate to what extent our observed frequencies differ from the frequencies expected provived that voters' preferences and their sex are independent. Such frequencies are called *expected* ones.

Expected frequencies are computed in the following way (recall the example on independent random variables above). First, we should calculate marginal frequencies getting a sum of frequencies by rows and columns (they are in blue):

	Male	Female	**Total**
Candidate A	40	25	65
Candidate B	10	55	65
Total	50	80	130

Then, multiply marginal frequencies for each pair of values (values here are A and B, *male* and *female*) and divide by a total number of people participated in the survey.

Thus, for the cell `[Candidate A, Male]` we have: $\frac{65 \times 50}{130} = 25$

For the cell `[Candidate A, Female]` we have: $\frac{65 \times 80}{130} = 40$

For the cell `[Candidate B, Male]` we have: $\frac{65 \times 50}{130} = 25$

For the cell `[Candidate B, Female]` we have: $\frac{65 \times 80}{130} = 40$

Now we can write our observed frequencies and expected frequencies side by side:

	Male	Female
Candidate A	40 25	25 40
Candidate B	10 25	55 40

Now our goal is to compare these frequencies, get a single measure based on such comparisons and decide whether this measure is large enough. If it is large, the observed differences are very different from ones that are expected in case of independence, so we can conclude that voters' preferences do not depend on their sex.

This single measure looks as follows (let us call it D):

$$D = \frac{(40-25)^2}{25} + \frac{(25-40)^2}{40} + \frac{(10-25)^2}{25} + \frac{(55-40)^2}{40} = 29.25$$

It is not necessary to remember this formula and to be able to compute everything by hand, but the very idea is important: we got a measure of divergence that does not take negative values. This measure D has a χ^2-distribution (*chi-squared*) with some number of degrees of freedom (here it is 2). And based on this value D we will calculate a p-value for a corresponding statistical test.

A chi-squared test (χ^2-test)

This test is used to decide whether two nominal (categorical, qualitiative) variables are associated or dependent. In some cases it is possible to detect which nominal variable is dependent. Consider an example above: it is clear that voters' preferences are dependent on people's sex, not vice versa.

Hypotheses:

- H_0 : two nominal variables are independent;
- H_1 : two nominal variables are not independent, they are associated.

P-value is computed as follows:

$$\text{p-value} = P(\chi^2 > D),$$

where D is the observed value of statistics that we calculated above on the data given, and χ^2 has a corresponding number of degrees of freedom (in case of 2×2 contingency table it equals 2).

Results:

- p-value $> \alpha \Rightarrow H_0$ is not rejected on the data given, so variables are independent;
- p-value $< \alpha \Rightarrow H_0$ is rejected on the data given, so variables are not independent, there is association between two nominal variables.