Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 7. Confidence intervals. (17 January)
Alla Tambovtseva

Interval estimates: an idea

Imagine that we want to know the proportion of people in a large city who suffer from depression. We randomly choose 1000 respondents, investigate them and see that 60% of people suffer from depression. So, we can report the proportion equal to 0.6. Should we be content with such an estimate? On the one hand, yes, our sample is large enough and representative. On the other hand, no matter how large our sample is, we get approximate results, an error of approximation is usually inherited in our conclusions. How to take this error into account? We can state how confident we are in our estimates and report the interval where the population proportion is likely to lie. Thus, we can calculate a*confidence interval.*

Confidence intervals: construction

We have a population proportion p that we do not know, but want to estimate. We estimate it based on a sample and do it with some error. While estimating a proportion based on the sample ($\hat{p}$) we allow the divergence from the true proportion p to be no greater than some value that is called *a margin of error.* So, we get:

$$\hat{p} \pm \text{margin of error}$$

$$\hat{p} - \text{margin of error} < p < \hat{p} + \text{margin of error}.$$

A margin of error depends on the level of confidence we choose. Usually in the social sciences we choose 95% confidence level or sometimes 90% and 99%. What does this confidence level mean? The rate of confidence in our results we get based on a sample. For example, if we independently repeat the same research on samples of the same size many times, 95% of intervals will include the true value of a proportion.

Suppose we decided to calculate a 95% confidence interval for the proportion of people in a large city suffering from depression. And we got a margin of error equal to 0.07. It means that if we independently repeat the research 100 times, in 95 cases the proportion estimated on samples will differ from the true population proportion no more than by 0.07 (provided that a margin of error does not differ from sample to sample).
How a margin of error is computed?

$$\text{margin of error} = const \cdot se,$$

where $const$ is some number depending on a confidence level and se is a standard error, a standard deviation of an estimate computed on a sample.

Recall: as the central limit theorem states, the distribution of a sample mean is $N(\mu, \frac{\sigma}{\sqrt{n}})$ *where* μ *is the population mean,* σ *is the population standard deviation and* n *is the sample size.*

The distribution of a sample proportion p *is* $N(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}})$.

In practice we never know a population standard deviation σ or a population proportion p, so we can approximate them using a sample standard deviation s and a sample proportion $\hat{p}$ correspondingly. Thus, a standard error for a sample mean:

$$se = \frac{s}{\sqrt{n}}.$$

And for a sample proportion:

$$se = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}.$$

During our course we will not compute standard errors by hand as well as constants for each confidence level, but we have to understand how a confidence interval changes when we change a sample size or a level of confidence.

Confidence intervals: properties

- **What will happen if we increase a sample size?**

  If we look at the margin of error, we see that it decreases as n becomes higher as it stands in the denominator of se.

  Thus, the higher is the sample size, the narrower the confidence interval is.

  This fact coincides with our intuition: the higher is the sample size, the more accurate results we get. So they are less dispersed around the true population parameter.

- **What will happen if we increase a confidence level?**

  Intuitively we can think that the more is the confidence level, the better it is. Let us consider the following example so as to make sure that this idea is wrong. Some candidate A asks us to estimate the proportion of people willing to vote for him in the forthcoming elections. We conduct a survey and report the following results: we are 100% confident that the percent of support is between 40% and 70%. Will the candidate be content with such results? He might believe us, but surely he will need something more accurate, more precise. So, we can decrease our level of confidence, but increase the precision. We can say that we are 90% confident that the percent of support is between 47% and 63% (approximately).

  So, usually we search for a trade-off between the confidence in our data and the precision of results. That is why, in research we usually take 95% confidence level, not 100%, since it is better to get more concrete results and be less confident (that is safe and realistic) rather than to get vague results and be absolutely confident (that is usually impossible in statistics).

Thus, the more is the confidence level, the wider the confidence interval is.

How to calculate the length of a confidence interval? To do it we should subtract the lower endpoint from the upper one:

$$Length(CI) = \text{Upper point} - \text{Lower point}$$

Now let us use the formulas discussed above:

$$Length(CI) = \text{estimate} + \text{margin of error} - (\text{estimate} - \text{margin of error}) =$$
$$= 2 \cdot \text{margin of error} = 2 \cdot const \cdot se = 2 \cdot const \cdot \frac{sd}{\sqrt{n}}.$$

Thus, we can derive two more facts:

- The length of a confidence interval descreases $\sqrt{N}$ times if we increase the sample size N times.
- The confidence interval is symmetric with respect to the population parameter.

Confidence intervals: interpretation

Suppose we know that the 95% confidence interval for the proportion of people suffering from depression is [0.55; 0.65].

✓ With 95% confidence we can say that the proportion of people suffering from depression lies between 0.55 and 0.65. If we independently repeat the same research on samples of the same size many times, 95% of confidence intervals will include the true proportion of people suffering from depression.

✓ If we independently repeat the same research on samples of the same size many times, in 95% of cases the true proportion of people suffering from depression will lie between 0.55 and 0.65 (assuming that the margin of error does not change from sample to sample).

✗ With the probability 0.95 the true proportion of people suffering from depression lies in the interval from 0.55 to 0.65.

? If we independently repeat the same research on samples of the same size many times, in 95% of cases the true proportion of people suffering from depression will lie between 0.55 and 0.65.

Why are the first two statements correct?

These statements contain the universal interpretation of a confidence interval and the explanation what a confidence level is.

Why is the third statement incorrect?

The level of confidence is not the same as the probability. When we say that we are 95% confident that we will pass the exam without preparation, we, possibly, know from the previous experience that in 19 cases out of 20 we passed exams without preparing. However, if we evaluate the probability of passing after taking the exam, it will be 1 (we actually passed) or 0 (we failed). The same situation is with confidence intervals. We calculate a confidence interval based on only one sample. And the true proportion of people will either fall in the interval (probability of falling is 1) or not (probability of falling is 0).

NB: There are different approaches to the probability. So, some researchers find such interpretations (with the word “probability”) correct. However, just to avoid ambiguity it is better to say about confidence or certainty, not probability.

Why can the last statement be ambiguos?

When we talk about the 95% confidence level, we imply the following: if we independently repeat the same research on samples of the same size, in 95% of cases we will get estimates of the proportion that differ from the true population proportion no more than by $1.96 \cdot se$ (recall three-sigma rule). And the value of se is different for each sample! One researcher can take one sample of 100 people and find that 50% of respondents suffer from depression ($se \approx 0.05$), another will take another sample of 100 people and find that 72% of respondents suffer from depression ($se \approx 0.04$), and so on. Thus, each person will get his own confidence interval, different from other intervals, but in 95 cases out of 100 these different confidence intervals will include the true value of proportion of people suffering from depression.

NB: It is usually assumed that a margin of error (a standard error) does not change from sample to sample. In this case this statement *contradicts to nothing*.

So as to make sure, you can look at this visualisation.