Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 1. Data collection. (01 November)
Alla Tambovtseva

Data collection. Population and samples.

In statistics and in data analysis there are two importants terms: a population and a sample. A **population** captures all objects of interest. A **sample** includes objects of interest that we directly investigate. Usually there are no resources (time, money, people) to investigate the whole population. Thus, we can work with a sample taken from a population so as to make conclusions about all objects of interest.

Question 1. *What problems do we face using such an approach?*

Studying a sample instead of the whole population might cause a serious problem concerning the adequacy of results if our sample does not correspond to the population well. As we are interested in correct results, we want to take samples that reflect the characteristics of a population fairly or correctly. In statistics and in quantitative research such samples are called **representative**.

Question 2. *If we know that in a particular city females take up 60% and males take up 40%, can we consider as representative a sample with 80% males and 20% females?*

If a sample is drastically different from a population, this sample is called **biased**. There are different types of biases (some of them we will discuss later), but most of them are highly dependent on the data collection process. If we conduct a survey among our close friends and report the results of this survey, the results will not be reliable since our sample is biased – rarely close friends can represent a population of inhabitants of a whole city in a correct way. The same problem will arise if we ask only females or only people older than 40 years in a pre-election survey.

Question 3. *What concept from the probability theory can we use to describe a population?*

Population can be modelled as a **random variable** with certain parameters. So, when we take a sample from a population, it can be considered as a sample from a random variable with a definite distribution. In a more formal way, when we take a sample of the size n, we capture n independent realizations of a random variable. To make it clear, let's look at the following example. A girl conducts an experiment: throws a dice and if she gets the score 6, she writes out 1, otherwise, she writes out 0. The result of this experiment can be described by a binary variable with the probability of success $p = 1/6$. What do we have if we get a sample from the binary variable like this:

$$0, 1, 0, 0, 0?$$

Indeed, we have records of throwing a dice 5 times. In other words, we ask the girl to throw a dice 5 times independently (i.e. redardless the results in the previous throws), check the scores and write 1's and 0's according to the values she gets each time.

Example 1. Consider the following situation: there is a population of people with different political views and we are interested whether each person supports liberal ideas. We can label liberals as 1 and non-liberals as 0. Then, we will get a sample of zeroes and ones taken from a binary variable with a parameter p, where p is the probability of a success, so the probability of meeting a liberal.

Example 2. We still study people's political views, but now we count the number of liberals in several groups of equal size n. In other words, for each group we count the number of successes in a series of independent experiments, where one experiment includes asking one person whether he is liberal or not. The number of successes is described by a binomial variable with parametrs n (number of experiments) and p (the probability of success in one experiment). Thus, asking m groups of people of the same size and with the same chance of meeting a liberal and counting liberals we will get a sample of the size m from a binomial variable with fixed n and p.

Example 3. It is a known fact proved by biological studies that people's height is distributed normally with some expected value μ and variance σ^2. It is intuitively true: there are a lot of people whose height is not far from the average and few people who are extremely tall or short. We can take 100 people (only males or females since the average height of men and women differs as well as the variance of the height) and get a sample from a normal random variable with $N(\mu, \sigma^2)$.

Now let's go on and solve some problems.

Problem 1. Consider a binary variable with the parameter $p = 1/4$. Suggest a representative (highly probable) sample of size $n = 10$ from this variable.

Solution. If we want a sample to reflect the features of a population (a random variable) well, such a sample should contain 1/4 of ones and 3/4 of zeroes. At the first glance, it seems to be impossible since $10 \cdot 1/4$ is not an integer, but it should not confuse us. As our sample is very small, the proportions of 1's and 0's in a sample can hold approximately, so it is okay to have 2 ones in a sample (1/5 of 10) or 3 ones in a sample (3/10 of 10). Hence, we can suggest the following sample:

$$0,\ 1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.$$

On the contrary, the following sample cannot be representative:

$$0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.$$

Problem 2. Can the following sample be a representative sample from a standard normal variable:

$$-8,\ -3,\ 4,\ -5,\ 7,\ 9,\ 2.5?$$

Solution. A standard normal variable is $Z \sim N(0, 1)$. From a three-sigma rule [1] we can derive the following: 99.7% of values of Z belong to the interval $[-3; 3]$. Let's look at our sample. Only two values fall in this interval! And these values are at the margins of the typical values of Z (2.5 close to 3 and -3 is the lower bound itself). Thus, we can conclude that this sample cannot be a representative sample from a population described by a standard normal variable.

[1] Look if do not remember what it is.