Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 2. Data collection. (08 November)
Alla Tambovtseva

This lecture is based on the Theme 2 from M.Sternstein. AP Statistics. Barron's. 2015.

Data collection

Methods of data collection:

- Census
- Sample survey
- Observational study
- Experiment

Now let's discuss each method in more detail.

Census

A *census* is an enumeration of all members of a population. Unlike governmental services, due to the lack of resources, a common researcher cannot collect data about every person in a country or at least of a town or a city. However, he or she has an opportunity to use the official statistics published by government. Often such an approach is hindered by the aggregated form of data since the level of aggregation is too high to provide the correct information about smaller units. These difficulties should not discourage us: further we will see that investigating a representative sample instead of the whole population is still powerful due to statistical laws. Besides, this method of data collection can be possible in the study of a particular community in the context of collective choice. For instance, we can have all roll call votes of the members of a legislative body (still anonymous due to the secret vote, but not aggregated) and study how these people make their decisions, whether their opinions on certain problems coincide, etc.

Sample survey

A *sample survey* is a study of a part of a population. A 'survey' here does not necessarily imply asking questions and recording people's answers, it is used as a general term, as a synonym for a 'sample investigation'. There are different types of sampling:

- **Simple random sampling:** taking a sample in such a way that every object in a population has an equal chance to be selected. For instance, we can assign numbers from 1 to 1000 to all master students of a faculty, run the random number generator[1] 50 times (setting the domain from 1 to 1000), and make a sample of those 50 students whose numbers were chosen by the random number generator.

[1]To be more precise, the random number generator returns pseudo-random numbers, not purely random since the computer algorithms are fixed, not random. We will talk about it later while generating random samples in R.

- **Systematic sampling:** a procedure that includes arranging the population in a particular order, choosing a starting point at random and taking every k-th object in a list. As the order used for arranging data is not connected with the features investigated, we can get a correct random sample from the population.
- **Stratified sampling:** a procedure that includes dividing the population into several homogeneous groups (*strata* [2], i.e. groups of objects with similar characteristics and groups are mutually exclusive), taking random samples from all strata and combining them into one sample.
- **Cluster sampling:** a procedure that includes dividing the population into several heterogeneuos groups (*clusters*, i.e. groups of objects with different characteristics) and taking a random sample of clusters.
- **Multistage sampling:** a procedure that includes several steps at which various sampling techniques can be used.

Example 1. (Systematic sampling) We have a list of all HSE professors' emails and we want to take a sample of 80 professors. We arrange the email addresses in alphabetical order and take every 10th email. When the list ends, we start from the very beginning and repeat this procedure until 80 people are chosen.

Example 2. (Stratified sampling) People living in Moscow can be divided into several mutually exclusive groups (strata) based on their age. So, we can get 5 groups of people: less than 18 years, 18-24 years, 25-40 years, 41-60 years and greater than 60 years. It is clear that no person can be assigned to more than one group simultaneously. Than we take a random sample from every group, get 5 samples and mix them up.

Example 3. (Cluster sampling 1) All master students of a particular faculty are divided into groups depending on a master programme. Pick 4 groups at random from the list of all groups and get 4 clusters. Why are they called *clusters*? Compared to each other, these groups are mutually homogeneous (each group corresponds to a particular specialization), but inside they are diverse since there are people of different age, background, nationality, income, etc.

Example 4. (Cluster sampling 2) Divide all countries into groups by their geographical position: Africa, Asia, Central America, Eastern Europe, European Union, Middle East, North America, Oceania, South America, The Caribbean. Take randomly 3 regions, for instance, Africa, European Union, Middle East. You will get a random sample of clusters containing states of different types: stable and unstable democracies, rigid autocracies, republics and monarchies, well-developed and developing countries, etc. Besides, there still will be geographical dispersion as we will not concentrate on a concrete region.

Last time we discussed that in statistics we usually work with sample estimates rather than with exact population paramenters. A **sampling error** is inaccuracy that is inherent in every sample survey that stems from the fact that sample estimates are approximate values of population parameters. A sampling error is a common thing and it cannot be

[2] *Strata* is a plural form. The singular form for *strata* is *stratum*.

completely removed while making inference based on a sample. However, we can control it varying the sample size and our beliefs about the reliability of data. The more is the sample size, the less is the sampling error (ceteris paribus[3] and considering representative samples). A **bias**, unlike the sampling error, is not an expected thing in a well-planed sample survey, it is a negative disbalance that occurs when some objects of a population have more chances to be included in a sample than other. There are different types of bias. The following types of bias should be better regarded as the possible sources of bias rather than classes since these types are not mutually exclusive and do not have clearly defined distinctive features. Some types of bias usually go together, some might be special cases of others, etc. The most important thing is to predict the potential bias inherited in the data collection process and to find ways to level it.

- *Household bias.* Occurs when one type of respondent is overrepresented while others are underrepresented since we investigate groups of different sizes in the same way. For instance, we take only one person from every household and record his or her characteristics. The households that include one person are represented correctly, but households with more than one person are not covered properly.
- *Nonresponse bias.* It is sometimes called *participation bias.* Occurs when only members of a population with certain features are included in a sample while other members are out of reach. For example, we conduct a door-to-door survey in working hours. Participants that are at home are pensioners, unemployed people, women with small children, etc. We will not include working-age people or students that are at school or at university.
- *Quota sampling bias.* Occurs when a researcher on purpose tries to 'adjust' a sample to keep the proportions of different groups in a population, so leaving no chance for randomness. It is okay when a sample resembles a population, but it is incorrect if a researcher selects objects with specific features and includes them (and only them) in a sample.
- *Response bias.* Occurs when respondents systematically provide unfair or misleading answers for a survey, usually because of the sensitivity of questions. The cases of not responding also can be considered as the results of this type of bias. Examples of sensitive questions: questions about income, voting, health, attitude towards political leaders, etc.
- *Selection bias.* Arises when a certain group of population members is included in a sample based on a particular feature. For example, if we study Russian people's behaviour in social networks and conduct an opinion poll only in Facebook, this will lead to a selection bias since the users of a more popular network Vkontakte are not covered at all (and we select this sample in a convenient way: using the network we like and use frequently).
- *Undercoverage bias.* Takes place if some groups of a population are not represented in a sample. Tightly connected with the nonresponse and the selection bias. Typical examples: asking people about their political preferences on the Internet (people that

[3] All else equal or held constant (latin expression often used in data analysis and statistics).

rarely use it are out of the scope); leaving questionnaries on the car (only car drivers or car owners are included).

- *Voluntary response bias.* Occurs when a researcher allows people not to return a questionnaire filled in or publishes it online with the message like "Fill in if you want/if you have time...". If people have an opportunity not to participate, we can face the case when only active people with special background (level of education, motivation, working hours) will decide to answer.
- *Wording bias.* Takes place when the question is formulated incorrectly. A question might have dual interpretation, the word order can be misleading and confusing, words can be complex, etc. All these mistakes lead to answers that are different from real people's beliefs and thus, to the misrepresentation of a population.

Observational study

A study in which a researcher simply observes the situation and carry out measurements. Observers are not able to affect the situation, all they can do is to detect changes in values measured or find associations between variables. An observational study cannot be used to find cause-and-effect relationships. To do this we have to be sure that Z is caused by X, not by some Y that is hidden from our view but that affects both X and Z. Such Y is called a **confounding variable**, a variable that influences both the independent variable X (cause, effect, factor) and the dependent variable Z (result).

Experiment

Unlike observational study where we cannot influence anything, an *experiment* is a controlled study that depends on the conditions provided by a researcher. Usually to detect the effect of a certain factor all objects of interest are divided into two groups: a treatment group and a control group. A **treatment group** is a group of objects that are affected by the factor studied and a **control group** is a group of objects not influenced by the factor. While making up examples of experiments, we can think of medical experiments where one group of patients is given the certain pills (a treatment group) and another group of patients is given a placebo (a control group). Conducting an experiment is a complex procedure that requires a carefully planned design and fulfilment of several conditions. One of those conditions is **double-blinding** when both a person responsible for distributing the influensive factor and a participant do not know which group the participant belongs to.