AdTrans 2023-1-PL01-KA220-HED-000158917

ADTRANS
ADVANCED DATA ANALYSIS FOR TRANSPORTATION AND MOBILITY ACADEMIC TEACHERS

Sample selection

Kristyna Rybova

Co-funded by the European Union

POPULATION AND SAMPLE

- Before starting research, it is necessary to decide who or what I will study?
- Population – the group about which we want to make a conclusion
  - We can almost never study the entire population and everything that concerns it.

→ The necessity of choice

  - Setting the procedure for creating a sample of observed units
  - The decision about the population and sample is related to the chosen research method.

POPULATION AND SAMPLE

- What area of research is the development of sample creation procedures related to?

POPULATION AND SAMPLE

- What area of research is the development of sample creation procedures related to?
  - ***Pre-election polls*** – elections allow us to verify the accuracy of polls
  - In the USA, election polls have been conducted since the beginning of the 20th century
  - *Literary Digest*
    - Sending correspondence cards to people selected from the telephone directory and car owner list
    - Correct prediction of the outcome of the presidential elections of 1916, 1920, 1924, 1928 and 1932
    - 1936 – 10 million tickets sent out, 2 million returned – according to the tickets, A. Landon won against F. Roosevelt – did it work?

POPULATION AND SAMPLE

- What area of research is the development of sample creation procedures related to?
  - ***Pre-election polls*** – elections allow us to verify the accuracy of polls
  - In the USA, election polls have been conducted since the beginning of the 20th century
  - *Literary Digest*
    - Sending correspondence cards to people selected from the telephone directory and car owner list
    - Correct prediction of the outcome of the presidential elections of 1916, 1920, 1924, 1928 and 1932
    - 1936 – 10 million tickets sent out, 2 million returned – according to the tickets, A. Landon won against F. Roosevelt – a bad prediction! Why?

POPULATION AND SAMPLE

- What area of research is the development of sample creation procedures related to?
  - ***Pre-election polls*** – elections allow us to verify the accuracy of polls
  - In the USA, election polls have been conducted since the beginning of the 20th century
  - *Literary Digest*
    - Sending correspondence cards to people selected from the telephone directory and car owner list
    - Correct prediction of the outcome of the presidential elections of 1916, 1920, 1924, 1928 and 1932
    - 1936 – 10 million tickets sent out, 2 million returned – according to the tickets, A. Landon won against F. Roosevelt – a bad prediction! Why?
    - Based on the "selection support", above-average wealthy respondents were selected, the situation after the end of the Great Depression - the poor preferred Roosevelt's New Deal program with reforms to support and restore the economy

POPULATION AND SAMPLE

- What area of research is the development of sample creation procedures related to?
  - ***Pre-election polls*** – elections allow us to verify the accuracy of polls
  - *George Gallup* correctly predicted the outcome of the 1936 election
    - New selection method – quota selection – quota on income structure
    - *American Institute of Public Opinion* (Gallup Institute)
    - Correct guess of the results of 1936, 1940 and 1944
    - 1948 – Thomas Dewey's victory estimate over Harry Truman
    - Why?

POPULATION AND SAMPLE

- What area of research is the development of sample creation procedures related to?
  - ***Pre-election polls*** – elections allow us to verify the accuracy of polls
  - *George Gallup* correctly predicted the outcome of the 1936 election
    - New selection method – quota selection – quota on income structure
    - *American Institute of Public Opinion* (Gallup Institute)
    - Correct guess of the results of 1936, 1940 and 1944
    - 1948 – Thomas Dewey's victory estimate over Harry Truman
    - Why?
    - Election polls ended in early October with Truman's support steadily growing.
    - A large proportion of undecided voters – in the end, they voted for Truman more often
    - Non-representativeness of the selection – quotas were set according to data from the 1940 census – after World War II, migration to cities – the sample overestimated the number of rural residents and underestimated cities where there were more Democratic voters

POPULATION AND SAMPLE

- Population (base set)
  - A set of units for which we assume that the research conclusions are valid
  - Clearly defined groups of people
- Sample (selection file)
  - The group of units we are actually examining
  - Sample research
  - Cheaper
  - Faster
  - Often technically impossible to survey the entire population
- Exhaustive selection
  - Complete match between sample and population
  - All population units included in the sample
  - E.g. census, organization database

REPRESENTATIVE RESEARCH

- Results valid for the sample generalizable to the population
- Matching of the basic and sample sets - the sample structure must imitate the composition of the population as accurately as possible

WHAT TO CONSIDER WHEN CHOOSING?

- Population size
- Population composition (homogeneous/heterogeneous)
- Complexity of the problem under investigation
- The intended level of statistical probability
- What information about the population do we have available?
- Resources (economic side)

SELECTION TYPES

Representative selections		Non-representative selections
Probability selections	Non-probability selections	
Simple random selection	Quota withdrawal	Snowball technique
		Purposeful selection
		Selection by availability
		Selection of critical cases
		A selection of typical examples

RANDOM SELECTION

- Every element of the population has an equal chance of being selected
- Represents all known and unknown characteristics of the population
- Selection support
  - A complete list of all units in a given population
  - From it, randomly selected units for research
  - Problem?

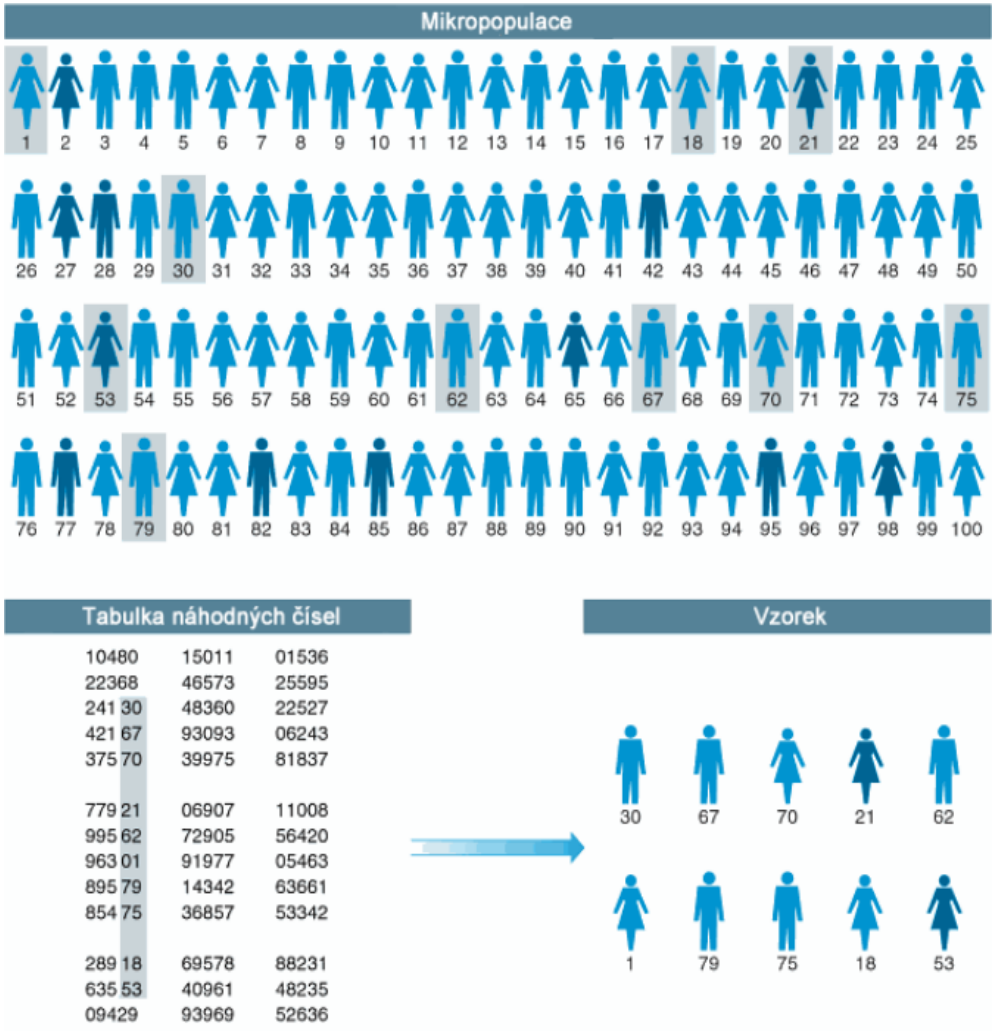
RANDOM SELECTION

- Every element of the population has an equal chance of being selected
- Represents all known and unknown characteristics of the population
- Selection support
  - A complete list of all units in a given population
  - From it, randomly selected units for research
  - Problem – Often non-existent or unavailable
- It is possible to estimate how the sample differs from the population (sampling error)
- Random selection ≠ whatever comes to mind at the researcher's fingertips
- Random selection ≠ who is passing by

TYPES OF RANDOM SELECTION

- Simple random selection not always effective
  - E.g. surveying secondary school students throughout the Czech Republic – the selection will usually include 1 student from a class/school → very time-, organizationally and financially demanding data collection
- Types of random selection
  - Simple random selection
  - Complex random selections
    - Clustering (group selection)
    - In cases where respondents are naturally gathered in groups, e.g. municipalities, schools, companies
    - First, randomly select a certain number of groups (e.g. schools) and then select only within the selected groups (e.g. students only in these schools)
    - Stratified random sampling
    - Division of the population into certain categories (strata), random selection within them
    - Different combinations of procedures

SIMPLE RANDOM SELECTION



https://is.muni.cz/el/1421/podzim2014/VIKBA12/um/50708019/7_Vyber_vzorku.pdf

It was or it wasn't, once upon a time there existed a small kingdom called Org. It was a kingdom where everything was well organized, and everyone was happy and content. Everyone, from soldiers to basic servicemen, complained about their pay, about food, about the preservation of ideas, about everything, and yet because everything was well organized, the government invited a foreign expert, Professor P. I. Toma, to conduct research on the attitudes in the army.

P. I. Toma arrived, constructed a questionnaire, and tested its validity. Because the kingdom was so small that there wasn't even a proper random number table there, he decided to use a systematic sampling technique for constructing the sample. The army of this small kingdom was so small, dignified, undersized, and consisted of a contingent of only 12,000 people. Professor P. I. Toma estimated that a sample of 200 people would provide an acceptable reliability interval and therefore chose step 60. He randomly selected the first unit. It was person number 31, and then he selected every sixtieth soldier. The research results were simply tremendous. No one had ever seen such a satisfied army. Everyone was happy in that small happy kingdom—until next spring, when the bloody uprising of basic servicemen began.

But you already know what happened: Simply in the kingdom of Org, everything was well organized. Even the members of the army were organized by platoons, with two officers in each platoon, then three non-commissioned officers, then the crew of basic servicemen, and each platoon had no more, no less than 30 people. Even our professor had bad luck because the chosen step coincided precisely with the double size of the platoon, but also because the first randomly selected person was an officer, and therefore every subsequent person also had to be an officer. Non-commissioned officers and basic servicemen were not included in the sample at all.

ADVANTAGES AND DISADVANTAGES OF RANDOM SELECTION

- Advantages
  - Controllability
  - Possibility of international comparison
- Disadvantages
  - High costs
  - Dependence on the selection support
  - Low return

QUOTA SELECTION

- Imitates known population characteristics in the sample structure
- Can only be used for populations about whose structure we have sufficient information
- Information sources for quota selection
  - Statistical information, especially the census
  - Other available overviews – about the structure of employees, students, etc.
- Quota regulation
  - Few characteristics
  - A selection of those that are clearly visible, identifiable, and the respondent is willing to list.
  - Regional characteristics - usually the size of the municipality, region
  - Sociodemographic characteristics - usually age, gender, education/occupation

SIMPLE QUOTA - EXAMPLE

Total respondents: 4

Sex:

Men 2

Women 2

Age:

18-34 1

35-49 2

50 and over 1

COMBINED QUOTA - EXAMPLE

Sex/ age	Men	Women	Total
18-34	1		1
35-49	1	1	2
50 and over		1	1
Total	2	2	4

ADVANTAGES AND DISADVANTAGES OF QUOTA SELECTION

- Advantages
  - Quick
  - Flexible
  - Anonymous questioning
  - Lower costs
- Disadvantages
  - Dependence on the timeliness of data for quota creation
  - Great influence of the interviewer
  - Difficult to control his work

PURPOSEFUL SELECTION

- Based on the researcher's judgment about what should be observed and what can be observed (visitors, audience, members of social groups, etc.)
- The researcher must precisely and clearly define the population to which the sample refers.
- It can be done as a representative

SURVEY

- So-called self-selection – not representative, problematic
- The selection of individuals is based on the respondent's decision to answer the survey
- The population to which the results apply cannot be defined.

SELECTION OF RESPONDENTS ON THE INTERNET

- Potentially very effective way of collecting data
- The problem – we often don't know how the sample was created
  - Self-selection
  - Only people with internet access
- Conducting representative research on the Internet is difficult
  - Research agencies usually have regularly rotating panels of registered respondents

SNOWBALL TECHNIQUE

- Not a representative selection
- Respondents indicate contact with other individuals
- Suitable for research on temporary populations (witnesses of events, participants of events, etc.)
- Often qualitative research

RETURNABILITY

- Not every selected potential respondent will be willing to participate in the research
- The reason for refusal may be related to some characteristics of the people addressed (e.g. diligence, respect for authority, openness, activity, as well as unintended barriers, such as lack of time, frequent travel away from home, etc.)
- Very successful research – return 60-70%, often more like 30-40%
- Returns can be influenced
  - by repeatedly addressing selected respondents,
  - the way the research is presented to potential respondents,
  - the public's interest in the topic,
  - (financial) reward to respondents for participation, …