Statistical file

A statistical file is a set of objects (data) that are the subject of a statistical investigation. If we have a complete set of data available (in descriptive statistics), we speak of a **basic set.** If this file is not available and we only have a selection from it, or if the subject of investigation is a random variable, we are talking about a so-called **sample file.** The number of elements of the set is called the **extent of the file.**

Basic file

- **The basic file** is specified by precisely defining its elements. These elements are determined either by enumeration or by establishing a clear rule (e.g. certain common properties), which is then the criterion for the element's belonging to the given basic set.
- Elements of the basic set can be various objects – persons, families, experimental animals, biological material, EEG recording, etc.
- E.g. the basic set can be a set of inhabitants of a certain territory in a given time period, a set of children with a certain congenital defect, a set of tissue samples from experimental animals, ...
- The **scope** of the basic set can be *finite* (e.g., demographic sets), or *infinite,* which is rather an ideal set, existing only hypothetically (e.g., the set of all possible results of experiments that can be performed in given experimental conditions or a set of all persons with a given disease).
- By a **homogeneous set** we mean **a set that is homogeneous,** i.e. in which all members have the same properties corresponding to a pre-selected criterion.

Selection file

Statistical inference methods can be used to draw conclusions about the basic set from the identified properties of part of its elements, the so-called **selection**. In order for this to be possible, it is necessary to choose a selection that is **representative**, i.e. a selection that reflects the properties of all the elements of the basic set in its composition. A selection that is not representative is referred to as **selective.** E.g. when determining the average height of boys in a population at age 10, a sample of boys who play basketball may be highly selective.

In order to quantify the degree of uncertainty that the selection is not representative, and thus support the scientificity of statistical inference conclusions, it is necessary to create a selection using the technique of *random or probabilistic selection*. These methods are to guarantee that every element of the base set has an equal chance of being selected, and every other element is selected independently of the ones we've already selected. Depending on the method of execution, we distinguish between different random selection techniques, e.g.:

- **Simple random selection**
  - is carried out by a suitable drawing technique. There are a number of drawing techniques, including the use of random number tables. However, the disadvantage of this procedure is the necessity of prior identification of elements (e.g. numbering), which is not feasible in practice for larger files.
- **Mechanical (systematic) selection**
  - it is conditioned by a certain, predetermined arrangement of the elements of the basic set. We include in the selection all elements that are separated by a certain selection step k, choosing the first element at random from the k elements at the beginning of the set.
- **Regional (stratified) selection**
  - is performed when the file can be divided into such areas that are internally homogeneous (they do not differ much in the observed characters) and heterogeneous among themselves (they can differ among themselves).
- **Group Selection**
  - is performed in cases where the base file is very numerous. Here we do not select individual elements, but groups of elements that form natural or artificial groupings (e.g. family). It is desirable that the groups are as large as possible and that the elements within the groups are diverse.
- **Multilevel selection**
  - is based on the existence of a certain hierarchical arrangement of elements of the basic set. We gradually get to these elements through higher selection units (e.g. cities – houses – households).

Links

- ws:Statistický soubor

Related articles

- Descriptive statistics
- Statistical inference
- Analytical study

References