STATISTICAL SAMPLING APPLIED TO HISTORICAL AND DEMOGRAPHIC SOURCES

Dr. Stanislaw BOROWSKI
(Poznan)

1. Representative investigations and statistical sampling are generally applied if the universe is strictly defined.

Besides this one of the following conditions must be met:

a/ there is no statistical source in the strict sense of the word,

b/ there exists a statistical source in the form of registration of all units and of the all variables to be investigated.

As an example of the first situation we can cite: various regional population investigations, interwar American unemployment investigations, British social surveys and others. Whereas the second situation takes place mostly during population censuses linked with representative research. [1/]

In the historic and demographic investigations this situation does not take place. There exists a statistical source biased in various ways. Let us look at the pecularities of the said source.

2. Statistical research of mass phenomenon takes place in three stages: a/ the observation and registration, b/ the reconstruction of the universe, c/ the inference and discovery of regularities. From these, observation and registration are necessary and sufficient to create a statistical source even if at the beginning they were not undertaken for statistical purposes.

Depending on the development stage of the society the mass phenomena either had not been at all subject of observation by man of the time, or were observed, registered and later became subject to further statistical research, or only the traces of past mass phenomena were observed and registered by man of later

generation. Thus there can be three forms of remnants of past mass phenomena: a/ undocumented in writings, b/ at least observed and documented in writings by man of the time, c/ observed and documented on the basis of remnants by man of later epochs.

In the first case we have to do with historical but not statistical sources. They can, however, be investigated only with help of statistical methods. I propose to call them inproper statistical sources.

In the second case the first stage of statistical research has already been passed. We have here to do with primary statistical sources.

The third case is wholly different. Totally different assessment criteria must be applied here. I propose to call these sources - secondary sources.

The further considerations are limited to primary sources.

From the moment of creation to the moment of exploitation, the statistical sources had their own history. The most important events in this history are all changes in the completeness and content of the sources. The knowledge of this is one of the basic elements in taking a decision on the proper use of statistical sampling.

3. Every historical and statistical source before worked upon must be properly assessed. We shall distinguish three kinds of assessment criteria: the formal criteria, the material ones and the criteria of the possibility of reconstructing the historical process. In the first group we may mention the following criteria: convergence of purposes of an observation by man of the epoch and the purposes of the present exploitation, immediateness of observation, homogeneity, adequacy and completeness of observation, representativity in choice of the units, types of selection, correctness of reconstruction and of inferences. The group of material criteria comprises the criteria of reliability. At last the group of criteria of possibility of reconstructing a historical process, for example of a historic and demographic process, comprises the following criteria: chronological and geographical order, identity of subsequent types of mass phenomena, the causal comparativeness.[2/]

Every of these criteria is being divided in subcriteria. Besides the ordinary assessment the criteria and subcriteria have to help in taking a decision on the proper use of statistical sampling.

4. Before working on the source the research worker must decide on whether to exploit it totally or partially.

If the source has been preserved completely all these methods of sampling can be used which are applied nowadays to a full registration of units and variables to be investigated.

The decision on a full or partial use of the source gets complicated if the historical statistical source from the beginning did not comprise all the units or has not been handed over in full. In that case on the basis of the history of source, aforementioned formal criteria and especially with a help of the criteria of completeness, representativity and types of selection, we must decide on the method.

5. The study of history of the source, especially of the causes and ways of appearance of shortcomings should establish: a/ whether before the creation of the source or appearance of shortcomings it was known which units of the universe not enter to the source, b/ whether between the factor which caused the shortcomings in the source during the creation or in the course of the further history and the studied variable there exists any relationship.

When both the questions are answered negatively, then should be established whether the number of units is sufficient. We shall avail ourselves here with the ordinary parameters used by fixing the sample size namely: confidence limits, standard deviation and per cent of significance.

If the number of units is sufficient then in the given here conditions we have to do with the quasi random sample and approximately representative. Such a sample will be treated as sample with replacement, without replacement or systematic, depending on the results of studies on the causes and ways of appearance of shortcomings in the source.

The characterized way of investigation must be repeated for every variable to be researched. The quasi random sample and approximately representative in respect to one or more variables to be researched, may be not representative as regards the others. It may happen that the sample in the historical source is not any random sample and representative for any variable.

In the demographic studies a sample in historical source is rather representative as regards biological and strictly demographic variables such as sex, age, number of children in a family, then for social variables such as vocation, education, income a. s. on.

When the sample in the source is not reprezentative in respect to some or even all variables we may deal with it in an arbitrary way, we can work on a total sample or the part of it only. The inferences typical for representative methods, however, are not applicable, the inferences must be limited to the investigated units only.

6. The application of statistical sampling is more difficult if the source had been created in many places. There is a similar situation if the shortcomings appeared during the preserving of the source in many places. We may mention here such historical and demographic sources as libri animarum, libri baptisatorum, copulatorum preserved in parishes, the first population censuses and others. In such situation the statistical sampling may be applied to in the following way:

a/ we shall separate the complete sources and the uncomplete sources in respect to every variable,

b/ we shall study whether a priori it was known in which places the sources will become complete, in which uncomplete,

c/ we shall study, whether between the factor which caused the uncompleteness of sources in some places and variable to be investigated there exists any relationship.

If the last two questions will be answered negatively we shall have to do with a random sample.

The complete sources from various places we may treat either as a connective sample or stratify them as in the actual representative investigations.

The problem of the sample size from every stratum must be solved in a simplified way. If the complete sources comprise a low per cent of units in relation to the universe we will deal with the strata as with the random samples from the strata. If the strata are very numerous we will establish the size of samples to be taken from either in proportion to the strata size according to the formula:

$$n_1 : n_2 \ldots n_p = N_1 : N_2 \ldots N_p$$

or according to the J. Neyman rule

$$n_1 : n_2 \ldots n_p = N_1 \sigma_1 : N_2 \sigma_2 \ldots N_p \sigma_p$$

The not representative samples in source as regards some variables may be treated in an arbitrary way. We can quit these samples at all, work upon either all units or some purposive units. The inference, however must be limited to these units only.

7. The purposive sampling is not considered here. In historical and demographic investigations we often deal with them. They are nearly always biased and therefore the inference typical for random samples can not be applied.

8. At last we will write down some words on two examples of application of statistical sampling in Polish historic and demographic investigations.

A very simple stratified sampling has been applied to libri baptisatorum, copulatorum and mortuorum from the end of the XVIIIth and the first half of the XIXth century for the region of Mazowsze. The complete certificate books were separated for various time periods. In every parish which had uncomplete birth, marriage or death certificates the research workers undertook studies on causes and ways of appearance of shortcomings in sources as well as on relationship between these causes and ways, and variables to be researched (birth, marriage, and death rates, sex, age, vocation, type of farm and other variables). It was stated that the most numerous shortcomings appeared only per chance. Because of small number of complete certificate books they were treated as stratified random samples.

The second example concerns archdioceses of Poznan and Gniezno. In 1848, Catholic archbishop L. Przyłuski got through a population census unknown to Prussian authorities. It was the first trial of census with inscription of names on Polish territories, but many methodological and organizational faults were committed. In many parishes the whole of population had been registered in respect to all variables in program of census. In other parishes, however, the Jews had been either wholly omitted or the total number of them was given only. In a part of parishes the census was limited to the Catholic only. Besides this in some parishes the age and vocation were not registered.

During the actual investigation the complete sources and every kind of uncomplete sources were treated separately. The studies on causes and ways of appearance of shortcomings were undertaken. The complete sources were stratified. In the first stage of investigation a proportional (12 %) random sample was taken

from every stratum. In further investigations the total strata were considered. The uncomplete sources were also taken into consideration but for control purposes only.

9. The application of statistical sampling to historic and demographic sources is different from actual investigation in two essential respects:

a/ In actual investigations the method of statistical sampling serves to the creation of the statistical source whereas in the historical investigations it is being applied to already existing sources.

b/ In the historical investigations the randomness should be established ex post in studies on the causes and ways of the appearance of shortcomings and on the relationship between the said causes and ways, and variables to be researched.

c/ As a result the principles of sampling are being treated in historical, statistical and demographic sources more liberally than in the actual investigations.

NOTES

1/ Current population Reports, Labour Force Series. C.A. MOSER, The Use of Sampling in Great Britain, Journal of the American Statistical Association, 1944; 44, 246.

P.G. GRAY, T. CORLETT, Sampling for the Social Survey, Journal of the Royal Statistical Society, 1950; CXII.

J. NEYMAN, On the Two Differrent Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection, Journal of the Royal Statistical Society 1934; 97.

C. GINI, L. GALVANI, Di una applicazione del metodo repprezentativo all' ultimo censimento italiano della popolazione, Annali statistica 1929, ser. IV, vol. IV.

2/ S. BOROWSKI, Charakter i klasyfikacja zródeł statystycznych, Studia Zródłoznawcze, vol. IX.

S. BOROWSKI, Kryteria oceny zródeł statystycznych, Studia Zródłoznawcze, vol. X.