From **Brunelli, Lina & Cicchitelli, Giuseppe (editors). Proceedings of the First Scientific Meeting (of the IASE).** Università di Perugia (Italy), 1994. Pages 409-415. Copyright holder: University of Perugia. Permission granted by Dipartmento di Scienze Statistiche to the IASE to make this book freely available on the Internet. This pdf file is from the IASE website at http://www.stat.auckland.nz/~iase/publications/proc1993. Copies of the complete Proceedings are available for 10 Euros from the ISI (International Statistical Institute). See http://isi.cbs.nl/sale-iase.htm for details.

WHAT IS THE MAIN GOAL OF STATISTICAL EDUCATION?

Vladislav V. Shvyrkov
National University, 536 Oasis Drive
Santa Rosa, CA 95407, USA

1. Introduction

Teaching broad aspects of statistical methods with the emphasis on computational statistics is important, but it is not the main goal of statistical education. The main goal of statistical education is to teach statistical methods in accordance with the principle problem of statistics.

The principle problem of all science is to trace laws and regularities of our lives on the basis of a sample. Statistics, as a universal methodological science, participates in the solution of this problem by analysing a sample with the help of quantitative methods. Everybody knows that a statistical analysis is only as good as the data on which it is based. Therefore, the analysis of the quality of data (sample) is a principle problem of statistical science. To enable students to resolve this problem, statistical education must concentrate its attention on the critical analysis of the unwarranted (untested) assumptions (i.e. homogeneity, representativeness, and randomness) of conventional statistics.

There is an ethical aspect associated with these unwarranted assumptions in the context of statistical education. Students have the right to know when and which assumptions are unwarranted in statistical research. The neglect of such issues provides a major source of the statistical abuses perpetrated by researchers and educationists.

2. Unwarranted assumptions

Every science is built upon a few axioms, or general truths. They cannot be logically proved, but we accept them as obviously true to actual fact.

Conventional statistics is not a science since it is not built upon axioms. It is based on the following three principal unwarranted assumptions; homogeneity, representativeness, and randomness.

The first unwarranted assumption is that a sample is homogeneous. This means that we deal with good data, which is very important since our decisions are only as good as the data on which they are based. Unfortunately, statisticians do not always know what good data mean, and how that 'goodness' may be tested.

The second unwarranted assumption is that a sample is representative of the population with respect to their standard deviations. This means that the dispersion of the homogeneous sample is approximately equal to the dispersion of the population. This assumption is not, and cannot be, tested because, according to conventional statistics, a population is not known *a priori*.

The third assumption is that the expected deviations from the linear stochastic regression model (population) are a random variable comprised of unobserved residuals. This randomness assumption is attempted on *a priori* grounds because there is no formal test for unobserved residuals. The examination of the observed residuals does not necessarily help in testing the randomness assumption for the following two reasons. Firstly, the observed residuals may be affected by mis-specification of the model. Secondly, the test for normality of the observed residuals is not a test for analysing the shape of the observed distribution. It is a test which is used to detect whether or not the residuals come from the normal population.

To test the above unwarranted assumptions, statistics must be built upon four axioms (Shvyrkov,1992, 1993).

The purpose of this paper is to build a statistical science on the basis of axioms in order to test the unwarranted assumptions of conventional statistics.

3. The principal problem of statistics

The concept of the quality of a sample in statistical terms is related to the concept of the sample's representativeness with respect to a homogeneous population. How do statisticians cope with the homogeneity problem? How do they define the homogeneity of the data set? The philosophy of conventional statistics is based on the assumption that the homogeneity problem is not a statistical problem, and that "the homogeneity of the material studied cannot be ascertained in principle" (Sachs, 1982, p. 83). Therefore, in statistics, the problem of the homogeneity of the data set is treated as only a subject matter problem.

In fact, it is difficult to overestimate the importance of the homogeneity problem. Statisticians of many different schools are aware of it. For example, Chuprov (1959) in 1910 wrote, "the move from relative frequencies to mathematical probabilities can be carried out only on the basis of a homogeneous data set". Boldrini (1972, p. 162) also underlines the importance of homogeneity in statistics. He wrote, "... an experimental probability is an induction; it is based on homogeneous factual data... ".

Referring to the quality of statistical information, Fisher (see Fisher, 1959) wrote, "The statistician is no longer an alchemist expected to produce gold from any worthless material offered him". The potential consequence of employing non-homogeneous data is the complete invalidation of statistical inferences based on them. Therefore, this author entirely agrees with Pearson and Hartley (1954, p. 83) that "it is the function of statistical methods to emphasise that precise conclusions cannot be drawn from inadequate data".

Unfortunately, statisticians do not have statistical methods for testing the data set homogeneity, and consequently they are in the position of having to produce gold from statistical data of unknown quality. The bottom line must be: our statistical inferences are only as good as the data on which they are based.

The solution of the homogeneity problem lies in probability theory. The theory of probability is the foundation of statistics. It traces effect from cause by employing the homogeneity principle. Regarding this principle, Khinchin (1961) wrote: "The homogeneity principle in probability theory is a fundamental and unique doctrine employed for theoretical predictions of probabilistic events in specific cases".

The homogeneity principle in probability theory is associated with the notion of the homogeneous population. The population is considered to be homogeneous if its events have equal chances of occurring. Hence, the main characteristics of the homogeneous population is the constant success probability of its elements.

Probability theory employs the deductive method of thinking, going from cause to effect, i.e. from a homogeneous population to samples. However, these samples may be representative or non-representative of the homogeneous population. The move from effect to cause is possible only if the sample is representative of the homogeneous population. This move is built upon inductive reasoning, formulated in terms of probability theory. The first requirement of inductive reasoning is that the population is homogeneous. The second requirement is that the sample is representative.

Inductive reasoning is the essence of the New Quality Philosophy in statistics (NQP). The NQP studies inward manifestations of homogeneous mass phenomena. It considers statistics to be a science and defines it as follows: statistics is a universal methodological science which quantifies the cause-effect relationship. This quantification is valid only when the sample is representative of the homogeneous population. According to this definition, statistical research must consists of two stages: firstly the evaluation of the sample quality, and secondly, the construction of the statistical model.

4. Axioms of the New Quality Philosophy of statistics (NQP)

The New Quality Philosophy of statistics (NQP) is built upon four axioms which are formulated in terms of probability theory.

The First Axiom

The homogeneous population is known *a priori* as a homogeneous invisible population (HIP). It consists of two groups of homogeneous invisible causes, successful (S) and unsuccessful (U). The successful causes exert a positive influence on the event. The unsuccessful causes exert a negative influence on the event. These two groups of invisible causes are homogeneous if two conditions are met. Firstly, each cause must have an equal chance of occurring. Secondly, each group of causes has its own constant influence on the effect.

Thus, the first axiom is that the HIP is a Bernoulli probability distribution in terms of homogeneous invisible causes: successful and unsuccessful.

The Second Axiom

The success probability (P) of the HIP is the proportion of the number of successful causes in the total number of all causes. It is considered to be an explanatory factor. Under the influence of this factor, the average effect is generated. The visible average effect is a mean of the statistical population which has to be representative of the HIP.

Thus, the second axiom is as follows: the mean of the HIP is equal to the mean of the representative statistical population.

The Third Axiom

The average effect, generated under the influence of the homogeneous invisible causes (HIC), is presented by the mean of the representative statistical population. The individual effect, generated under the influence of the HIC, is presented by a statistical datum. The statistical datum is less than the mean under the influence of unsuccessful causes. It is greater than the mean under the influence of successful causes, and it is equal to the mean under the influence of both causes, unsuccessful and successful. Thus, combinations between the HICs determine the size of a statistical datum with respect to the mean of the statistical population when it is representative of the HIP. These combinations are the following:

$$U \cap U = x_1 < \bar{X}$$
$$S \cap S = x_2 > \bar{X}$$
$$U \cap U = x_3 = \bar{X}$$
$$S \cap U = x_4 = \bar{X}$$

Therefore, a statistical datum can be considered as a combined event or as an indivisible sample with the corresponding probability. Probabilities of combined events can be computed as soon as the success probability of the HIP is known. This means that each datum has an individual probability or, according to Markov's terminology, "a single case" is associated with its probability (Markov, 1911).

Thus, the third axiom is that a statistical datum is a combined event, or an indivisible sample, originated under the influence of the HIC.

The Fourth Axiom

The HIP is characterised by the variable, consisting of two types of homogeneous invisible causes, and by the success probability (P). The set of homogeneous invisible samples (HIS), drawn from the HIP, is characterised by the variable, consisting of combinations between the HIC, and by probabilities of this variable.

The size of combinations between the HIC must be equal to two in order to meet requirements of the principle of minimum and the principle of representativeness. The requirement of the principle of minimum is met when the size of the HIS is equal to the number of causes of the HIP, that is, two. The requirement of the principle of representativeness with respect to the HIP is met when the number of representative combinations between the HIC is equal to 50% or more.

For example, we expect the following combinations between the HICs when homogeneous invisible samples are drawn from the HIP:

$$U \cap U, S \cap S, U \cap S, S \cap U$$

These combinations are combined events (CE) or the HIS. The set of the CE consists of two non-representative combined events (NCE)

$$U \cap U, S \cap S,$$

and two representative combined events (RCE)

$$U \cap S, S \cap U.$$

Let us designate the NCE by 0 and the RCE by 1. Then our set of the CE in terms of the theoretical values of representativeness (TVR) can be written as follows:

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

This set of TVR (STVR) is considered to be a representative set of the CE (RSCE) with respect to the HIP because it contains 50% of the RCE.

Our STVR is a variable of the representative set of combined events. Probabilities of this variable are computed with the help of the multiplication formula for independent events (causes) when the success probability of the HIP is known. This probability distribution of the TVP is representative of the HIP because two conditions are met:

1. The variable of the CE is a representative set of the HIS.

2. The success probability of the CE is constant. It is used for computing probabilities of the variable of the CE. This means that the probability distribution of the CE is generate under the influence of the explanatory factor, which is a success probability of the HIP.

Since each statistical datum is a combined event, the fourth axiom can be formulated as follows: the data set is good if it is a representative set of the combined events generated by the HIC.

Bibliography

Boldrini M. (1972), *Scientific Truth and Statistical Method*, Hafner Publishing Co. New York.

Chuprov A.A. (1959), *Essays on the Statistics Theory*, Gostatizdat, Moscow (in Russian).

Fisher R.A. (1959), *Statistical Methods and Scientific Inference*, Oliver and Boyd Ltd., Edinburgh.

Khinchin A. (1961), *The Frequency Theory of R. von Mises and Modern Ideas of Probability Theory*, Problem of Philosophy, #1, pp. 91-102; #2, pp. 77-89 (in Russian).

Markov A.A. (1911), On the Basic Principles of the Calculus of Probability and on the Law of Large Numbers, in *The Correspondence Between A.A. Markov and A.A. Chuprov on the Theory of Probability*, Springer-Verlag, New York (1981).

Pearson E.S. and Hartley H.S. (1954), *Biometrika Tables*, Vol. 1, 3rd ed., The University Press, Cambridge.
Sachs L. (1982), *Applied Statistics. A Handbook of Techniques*, Springer-Verlag, New York.
Shvyrkov V.V. (1992), *New Thinking about Quality Control*, Throwkoff Press, Santa Rosa, CA.
Shvyrkov V.V. (1993), *Axiomatic Principles of Statistical Science*, Throwkoff Press, Santa Rosa, CA .