What is Missing Sometimes to Enable Statistical Methods to Increase Their Cognitive Capacity?

Assoc. Prof. Nicolay Stoenchev, Ph.D.

Summary: The article considers the most common made subjective errors in applying statistical methods for analysis of economic processes and phenomena. Thesis advocates the need for a thorough qualitative analysis before proceeding to the processing of empirical data using specialized statistical software. It`s highlighted the importance of knowing the applying method in depth and with aim to obtain accurate results to avoid the mechanical custom approach. It is suggested for the combat with subjective errors team method to be used as a working principle and during the formation of research teams to attract qualified statisticians as participants or consultants. In this aspect, are achieved good practices, but they need to find wider application.

Key words: statistical methods, statistical analysis, subjective errors.

JEL: C14.

Introduction

The modern version of the Statistics Act and its supporting normative documents has created a favorable environment for the development of statistical activity in Bulgaria The new methodologies, which are used by the National Statistical Institute, and its active publishing and distribution activity have made statistical information still more reliable, comparable with other EU states and easily accessible for users of any rank – from the state administrative bodies and regional administrative structures to all citizens and companies who show interest. Not a small part of the information searched can be downloaded directly from the web-site of the department which is structured rationally, can be approached quickly and suggests references to the main data bases. These facilities and the growing number of users, who have access to Internet, has enhanced incredibly the prestige of statistical information and aroused wish such information to be used almost everywhere and in everything when a definite thesis must be substantiated. The temptation of using statistical methods of analysis at random and hastily has intensified, because this is prestigious and calculations can be made automatically with the help of functions embedded in the office product Excel or by specialized statistical software products. The technical ease of the processing of information has created a false impression that the essence of the statistical methods and their requirements in respect with the available information need not be known profoundly. It is relied on the fact the manufacturer, who developed the software, has predicted everything. But we should not forget that the statistical functions which are embedded in the standard office packages and the special statistical software programs which

are used on a large-scale still do not possess the capabilities of artificial intelligence and any researcher /student, scholar or manager/ is expected to do creative thinking activity. The application of a roughly empirical or, to put it in other words, overexposed consumer approach hides in itself a possibility for damage both to the consumer, who faces the risk of obtaining results, having nothing in common with the actual reality, and to statistics as a science which methods are discredited undeservedly.

These are in our opinion some of the reasons for the frequent cases of using statistical methods in an unprofessional way – 'just for illustration' or 'in support of a desired thesis', as we exclude the possibility of intentional manipulation which would certainly bring about wrongful management decisions, based on pseudoscientific truths.

In fact, superficial and false interpretation of the statistical data is a phenomenon that has been known for a long time and that has found an expression in the anecdote about the cucumbers which is popular among statisticians[1].

The purpose of this article is: Not to denounce a definite personality or a scientific research, but to outline the shortcomings which are most frequently made, when statistical methods are used, and to minimize the risks to draw up untrue and inaccurate conclusions for subjective reasons.

That is why the examples which will be used are hypothetical and any resemblance to real persons and scientific research is purely accidental.

Exposition

As a result of the extensive observations, related to research projects, the reading of scientific products and the supervision of graduates, we can sum up the following basic issues in response to the question: 'How should statistical data not be used in the analysis of social and economic phenomena and processes?'

1. When relationships and dependencies are analyzed, the requirement for priority of the qualitative analysis is not observed. It is directly experimented what value will be obtained in the calculation of a correlation coefficient without the nature of the phenomena having been studied and without having been assessed competently from the point of view of the specific discipline area /biology, economy, demography, etc./ if it is logical to expect the existence of such a relationship. On principle it is useful to perform a check of the statistical hypothesis, before we proceed to measuring the power and the direction of the relationship.

2. When a suitable correlation coefficient is selected, the scale of presentation of the data is not taken into consideration or the data from a weak scale are conditionally transformed into a stronger one by means of an expert valuation, as numerical total ratings are assumed. Then such are processed by parametrical methods. In this way the lack of accurate information can very delicately be circumvented and a thesis which is desired and convenient to the researcher be proven, but this is too risky. If the obtained results are published, saving the truth that an expert valuation has been used in the algorithm, this might mislead the users.

3. When relationships and dependencies are analyzed on the basis of data presented in the form of dynamic lines, methods are used which are suitable for analysis of statistical aggregates, i.e. the existence of autocorrelation between the elements of the dynamic lines, which is

[1] Briefly, it says: 'As a result of a statistical research it has been found out that 99.99 % of the persons, who died last year, consumed cucumbers, which gives a reason to think that their consumption is not safe to the health...'

usually related to the availability of a tendency, is not checked in advance. In respect with this issue N. Velichkova has written: 'It has been found out that the correlation coefficients which are calculated on the basis of autocorelated lines most often exceed considerably the actual correlation coefficients'. The author has presented profoundly many methods for elimination of the autocorrelation in the economic temporary lines[2].

A similar problem exists also, when regression analysis is applied on the basis of temporary lines. In this case it is necessary to check the independence of the residual elements around the regression line. If the residual elements are autocorrelated, the model is not good and the check of the significance of its parameters and their confidence intervals cannot be identified by using the classical models of regression analysis[3].

For example: By applying a formal approach and without taking into account the above-mentioned considerations, it can be proven that the availability of a marketing strategy in the production enterprises is related directly to the growth of the Gross Domestic Product. The number or the relative share of the companies which have a written marketing strategy is studied for several consecutive years by a representative sampling. It turns out that this indicator /number or relative share/ grows continuously. This growth is compared to the indicator of GDP produced which in normal conditions will also grow. As a result of the regression analysis, it appears that at a growth of 1 % of the number of companies, which have such a strategic document, the gross product will grow, for example, by BGN 300 per capita of the population. The conclusion is rather optimistic, only if it were true. In this case the very nature of the factor is also subject to discussion. Even if there is such a strategy in the company, it is likely that it is not used or is wrongful and has a negative impact on the final financial results.

4. Other insignificant 'clever devices', which undermine the confidence in the statistical methods, are:

- Revealing a trend on the basis of short dynamical lines, consisting of three to four elements, without performing a subsequent check for statistical significance of the parameters of the models obtained;
- Interpolation of missing data for interim years in short dynamical lines with the help of an expert valuation, as after that the lines are treated as sufficiently long. For example, lines of three values /for 1990, 1995 and 2000/ are supplemented by means of interpolation and are accepted as lines of primary data with 11 values for the period 1990 – 2000. Then such lines are used for the purposes of correlation and regression analysis without checking for a trend[4] which might have been introduced artificially in the interpolation. The obtained results are convenient to the researcher because they indicate the presumed dependence in the desired intensive form and can stand all checks due to the sufficiently long interpolated dynamic line.

In the end, as an antithesis to the popular proverb: 'You can make an omelet without breaking the eggs' because the precision of the method, which produced wrong conclusions, is put under question.

[2] Velichkova, N., Statistical methods for study and prediction of the development of social economic phenomena, Sofia, 'Nauka i Iskustvo' Publishing House, 1981, p. 289.
[3] Velichkova, N., Statistical methods for study and prediction of the development of social economic phenomena, Sofia, 'Nauka i Iskustvo' Publishing House, 1981, p. 273.
[4] Usually, the existence of a developmental tendency is a reason for the existence of autocorrelation in the economic temporary lines.

5. Sometimes the condition is not taken into account that it is impermissible to include two or more factor variables, which are proven to be mutually related, in the one and the same regression model. This phenomenon is called colinearity, respectively multicolinearily, and has a negative impact on the research results[5]. The selection of factors to be included in a multifactor regression model is within the exclusive competence of the researcher. Usually there is a tendency of including maximum number of factors, although the quantity of the available information is limited. The possibility of studying the relationship through a series of single-factor models, which can be complicated stage by stage by addition of new variables, meanwhile measuring the achieved improvement of the model, is underestimated.

In other cases it is preferred to use general quantities as a factor, obtained by summing of several variables. This makes difficult the interpretation of the constructed models. For manufacturers of goods and services it would be useful, if the studied factors which influence the final economic result of their business are maximum simplified and specified in order to be able to achieve really the revealed effect in real conditions, by way of suitable management decisions which are expressed in influencing the factor variable or conforming to its variation.

6. When statistical graphic representations are used, it is not sometimes given enough attention to the determination of the scope and to the ratio between the lengths of the coordinate axes. The inappropriate ratio between the axes of the chart can modify the shape of the graphic image to such a degree that the process represented may seem much more intensive or slower than it is practically /for example, when a line chart is used, an unrealistically steep or much smoother graphic line can be obtained than it should actually be/.

The graphical method of presentation of the statistical information ensures that such information is quickly perceived visually, which can be used for advertising purposes or as an application to textual information of top managers. This requires elaborating the statistical graphic representation not merely as an impressive colored picture, but as a synthesized, correct and accurate expression of real facts.

7. It is considered pointless to disclose negative results and unconfirmed statistical hypotheses. According to us, the negative scientific result is valuable as well /for example, the lack of dependence between the analyzed phenomena/. It outlines the perimeter of a territory already studied which would save time and powers to anybody who is planning to follow the same path or would raise the question: What are the reasons for the lack of compliance of this result with the expectations? Insufficient information, incorrectly selected factor, unsuitable method of analysis, etc.?

8. When applied research is conducted, the significance of collection of qualitative information is underestimated, as it is mainly accented on the selection of complex and impressive methods of processing. For example, at the defense of a graduation paper, devoted to the study of factors that influence the development of tourism in a famous Bulgarian mountain resort, the she-graduate stated that she conducted a representative research among the local population on the conditions of development of tourist business in this town. In order to form the sampling she used random selection. To the question of the Examination Board: Which of the methods of random selection known was

[5] About the nature of the phenomenon, the harm from it and the ways of its elimination, see in greater detail: Saikova, Iv., Statistical analysis of relations and dependencies, 'Nauka i Izkustvo' Publishing House, Sofia, 1981, p. 348.

used by her?, the she-graduate answered: 'One day I stopped at the square of the town center and inquired anyone, who walked by, about the problems in tourism ..."

In this case the fact is neglected that no results can be expected from doubtful or inappropriate data, even after hard statistical processing, that are suitable for making of correct management decisions.

9. Sometimes the insufficient training of newly admitted assistants to a higher education facility on a teaching methodology results in their underestimating the didactical principles later on (from the simple to the complex, from the known to the unknown, etc.).

For example, sometimes a methodology for application of complex statistical methods is taught in a narrow discipline area (branch or activity) without the curriculum providing for studying a course in 'General Theory of Statistics'. Probably a person, who has extensive research experience, can explain well these methods, which are actually 'advanced flying in statistics'. But it remains unclear: What would be the use to a student who does not know the alphabet of statistics? (for example, the requirements which must be observed, when an arithmetic mean and its properties are calculated, knowing that it is involved in the realization of almost all methods).

Conclusion

1. It can confidently be stated that in order to avoid the above-mentioned imperfections in the use of statistical methods, a more serious significance should be given to the partnership between experts from various specialties and to team work. Especially when the study has an interdisciplinary nature and requires knowledge in statistics, as well as and in several other disciplines /agronomy, economy, sociology, computer information systems, etc/. The integration of a qualified statistician with the team or the attraction of such a specialist as a consultant would guarantee a much higher degree of reliability and correctness of the obtained results. This can minimize the publication of scientific results which arouse uncertainty in the private business representatives about the capabilities of statistical methods to reveal regularities and outline prospects that promote adequate strategic decision-making. There are already good practices of formation of relatively stable scientific teams to work by projects with similar topics, in which scholars from various disciplines, departments and institutions are involved. This approach is encouraged in the participation of competitions, based on national and international funding. In this way conditions are created for the minimization of subjective mistakes which cannot be measured or predicted, as meanwhile a more efficient utilization of the scientific potential and of the available research facilities is ensured.
2. When doctoral and post-graduate students study statistical methods alone in order to be able to use them in scientific research, it would be useful that they do it on the basis of specialized textbooks and training aids with a guaranteed high scientific level [1, 2, 3, 4, 5, 6 and other], not on the basis of other applied researches where the methods can be interpreted incompletely, inaccurately or incorrectly for various reasons. ИА