DE GRUYTER OPEN

Folia Oeconomica Stetinensia
DOI: 10.1515/foli-2015-0041

WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA
UNIWERSYTET SZCZECIŃSKI

STATISTICAL COMPUTING IN INFORMATION SOCIETY

Prof. Czesław Domański[1]

Alina Jędrzejczak, Ph.D., Associate Professor[2]

University of Lodz
Faculty of Economics and Sociology
Institute of Statistics and Demography
POW 3/5, 90-255 Łódź, Poland

Centre of Mathematical Statistics, Statistical Office in Łódź
Suwalska 29, 93-176 Łódź, Poland

[1] *e-mail: czedoman@uni.lodz.pl*
[2] *e-mail: jedrzej@uni.lodz.pl*

Received 20 May 2015, Accepted 3 December 2015

Abstract

In the presence of massive data coming with high heterogeneity we need to change our statistical thinking and statistical education in order to adapt both - classical statistics and software developments that address new challenges. Significant developments include open data, big data, data visualisation, and they are changing the nature of the evidence that is available, the ways in which it is presented and the skills needed for its interpretation. The amount of information is not the most important issue – the real challenge is the combination of the amount and the complexity of data. Moreover, a need arises to know how uncertain situations should be dealt with and what decisions should be taken when information is insufficient (which can also be observed for large datasets). In the paper we discuss the idea of computational statistics as a new approach to statistical teaching and we try to answer a question: how we can best prepare the next generation of statisticians.

Keywords: statistical education, statistical computing, big data, information society

JEL classification: C1, C4,C8, O3

Some people hate the word statistics whereas I find this word beautiful and interesting. If statistics is not used in a primitive way but we use more sophisticated statistical methods and interpret them carefully, then it turns out that statistics is extremely efficient in solving complex phenomena. Statistics is the only tool thanks to which it is possible to pave the way for all those who deal with the science of the human.

Francis Galton (1822–1911)

Introduction

The ability to compile and analyze data is of fundamental importance for all citizens so as to enable them to participate fully in professional, social and personal activities under conditions of information society.

It seems obvious that nowadays no statistical research can be done without substantial informatics support. However, the role of computational statistics in statistical education and discovery, including the big data analyses, has been under-recognised even by peer statisticians. Especially in the presence of massive data coming with more heterogeneity we need to change our statistical thinking in order to adapt classical statistics, still invaluable in the context of big data analysis, and software developments in statistics that address new challenges. The new challenges comprise numerous problems coming from data characteristics that are difficult to deal with: scale (being still serious computational challenge), "curse of dimensionality" (millions of predictors, heterogeneity), more real-time needed (streaming computations every few minutes/ hours), temporal and spatial correlations etc. Moreover, the data which come from two or more sources and are characterized by different stochastic mechanisms may sometimes be mixed and registered together, which leads to polluted samples. For all these cases the traditional stochastic models, suitable for primary events (the ones which actually occur) may not be appropriate for description of registered events (observed data) unless necessary modifications are introduced.

The adaptation process should be based on incorporating into curriculums the topics as: exploratory data analysis (EDA) and visualization, advanced statistical modelling and forecasting, hypothesis testing from randomized experiments, planning of adaptive experiments, to name only a few. They all need special teaching methods based on both statistical and computer-science tools. Similar approach can be utilized to make students familiar with computationally intensive statistical methods including resampling methods, Monte Carlo Markov chains (MCMC), local regression, kernel density estimation, artificial neural networks and generalized additive models.

The main objective of the paper is to introduce a concept of statistical computing, a new branch of statistical education addressing the abovementioned challenges, as the interface between statistics and computer science. Within this idea a broader concept of computing should be taught as a part of general statistical education.

1. Statistical computing

The general concept of statistical computing (or computational statistics) dates back to the times before the big data could even has been considered. It is worth mentioning that during the 41st Session of the International Statistical Institute (ISI) in 1977, the International Association for Statistical Computing (IASC) was founded as a Section of the ISI. The terms *computational statistics* and *statistical computing* are often used interchangeably, although Lauro (1996), a former president of the IASC, proposed making a distinction, defining statistical computing as "the application of computer science to statistics", and computational statistics as "aiming at the design of algorithm for implementing statistical methods on computers, including the ones unthinkable before the computer age (e.g. bootstrap, simulation), as well as to cope with analytically intractable problems".

More precisely, *statistical computing* can be understood as text processing, data processing, statistical calculations, knowledge and experience in data interpretation from the point of view of various applications; while *computational statistics* is concerned with analytical aspects of data software and implementation of statistical procedures with the use of appropriate programmes.

In the process of education we acquire and improve three basic abilities: reading, writing and counting. The information society requires its members to possess the fourth basic ability: the ability to reason inductively on the basis of insufficient premise. Therefore, a need arises to know how uncertain situations should be dealt with and what decisions should be taken when information is insufficient (what can be observed also for large datasets!). Statistical computing, as a section of mathematical statistics, would allow us to acquire and develop this kind of ability. The new programmes of computational statistics should be designed by statisticians and IT specialists; the cooperation of these disciplines seems necessary as for most advanced statistical problems the profound informatics technology knowledge is not sufficient (see Stefanowicz, 2001).

Lectures on statistical computing are delivered at numerous American universities and the subject curricula encompass, inter alia, the following problems:

- programming,
- data management,

- statistical data processing with the use of software for statistical analysis,
- graphic methods,
- simulation experiments,
- new statistical procedures,
- Monte Carlo markov chains,
- symbolic calculations.

Statistical computing also includes the analysis of logical binary data. The section of statistics called Logical Analysis of Data (LAD) has the following characteristics:

- binary data analysis in order to discover logical patterns, generating data classification rules,
- possibility of applying the discovered rules for any type of data,
- possibility of applying the discovered rules for various application problems (classification, imputation of missing data).

Nowadays statistics is developing as meta-science whose main subject is logic and methodology of other sciences – the logic of decision-making and the logic of experimenting. The future of statistics lies in a successful dissemination of statistical ideas among researchers representing different fields of science. The success will also depend on the way in which fundamental problems are formulated in other fields of knowledge. When the logical approach is taken then the methodology of statistics is likely to expand so as to include not only information provided by data but also expert opinions in the process of estimation and decision making under uncertainty. According to Rao (1994), the following relationship can be formulated as a good illustration of the statements mentioned above:

Uncertain knowledge + Knowledge on the extent of uncertainty = Useful knowledge

Bessant and MacPherson (2002) conclude that the methodological nature of the discipline of statistics sets it apart from mathematics while drawing it closer to a host of research-oriented fields. They also point out that the necessary reformulation of statistics curricula and pedagogy is closely aligned with the roots of the discipline, for example, "working with data" and applying statistical methods to diverse research problems. Similarly, Bryce et al. (2000) indicate that the courses of statistics should include topics dealing with the following "core" areas, comprising not only mathematical but also computational and managerial skills:

- *Skills in statistical science: mathematics based* – such as data collection, data analysis, correlation, and statistical theory (e.g., variability, probability, and confidence).

- *Skills in statistical science: non-mathematics based* – for example, communication, collaboration, and project management skills.
- *Computational skills* – pertaining to word-processing, data handling, and statistical computing.
- *Mathematical foundations* – including calculus and linear algebra.
- *Substantive area skills* – for example, knowledge, skills, and experiences gained in a minor concentration that facilitate the interpretation of statistics in an applied context.

2. Vision of statistics and statistical computing

Statisticians can work in various fields of science gaining at the same time a vast knowledge on different aspects of their own discipline. Living in the world of constantly growing specialization, statisticians may combine their general and specialist skills. As there is a high demand for this kind of unique combination, statistics can provide an excellent opportunity for a life-long career .Looking towards future we can state that the inclusive approach towards statistics and the influence it has on research, training and career prospects is the approach we should take in our future actions. Another question which arises here is the question of the essence of statistics or the fundamental shape it will take in the years to come.

It seems that the time has come when it has to be admitted that computer science should take the place next to mathematics (and probability) as they are essential components of statistics and the disciplines which form the basis of our research. The examples of statistical computing applications include databases and data management, algorithms, computational statistics, artificial intelligence and machine learning.

Moreover, incorporating selected problems of computer science into the system of education of prospective statisticians will provide a range of interdisciplinary possibilities. Let us examine two of such possibilities which emerge on the borderline of statistics and computer science that can be defined as statistical computing.

The first area to be examined is *software engineering*. Although it is relatively new, this area seems to be a critical one in the information era. Development and production of reliable, high quality and complex software have become critical issues on the global market of computer technology. The ability to develop computer software and organize production process effectively is of key importance for the future economic power, competitiveness, and national safety. Statisticians and engineers should establish an effective cooperation. One of the biggest challenges

here is the statistical control of the process of software development which has been achieved in a more traditional production process of computer equipment.

The second area which is now of prime importance, is the area of *data digging*. It could be perceived by a statistician as the analysis of large databases. However, a closer examination of the data mining process will reveal that it includes a lot of statistical components such as: statistical graphics and cluster analysis. Moreover, this process provides an excellent opportunity to use many statistical ideas- modelling, sampling, robust estimation, detection of non-standard observations, dimension reduction etc. On the other hand, some new problems emerge so it would be a good idea to acknowledge them and take an active part in solving them with the use of statistical computing.

Looking forward to the future, statistical curricula for graduates of the 21stcentury should be based on certain aspects of computer science, much in the same way as they are based today on mathematics. It is worth noting and emphasizing that some needs of IT specialists and statisticians do overlap. Therefore, it would be in the interest of both the communities to respond to these needs and the statistical computing, a new section of statistics, will certainly help in the process.

Discussion on data digging will lead us in a natural way to a new range of topics and trends in the field of statistical data. The type of challenges that we are faced with seem to be a good indicator of the personal and scientific profile which are to be adopted by statisticians of the beginning of the 21st century.

Let us now concentrate on the scale of the problems.

1. Surprisingly, one of the up-to date trends is the tendency towards small problems. For example, in some branches of industry the competition is so tough, time horizons are so short and data collecting is so expensive that we are forced to act under unfavorable conditions where plans and decisions have to be taken on the basis of smaller and smaller amount of data.
2. Obviously, the real pressure is placed on bigger and bigger datasets and statisticians well-trained in variability show a tendency to simulate this direction.
3. Moreover, in many application areas we can observe the amounts of data that are really amazing:
   – Gigabytes in telecommunications,
   – Terabytes of data on the global climate changes.

However, the amount of information is not the most important issue.The real challenge is the combination of the amount and complexity of data. The so-called big data problems are frequently unsuitable for standard statistical solutions to be used.

Statisticians have just begun to face the challenge of massive datasets when some potentially valuable approaches seem to emerge e.g.

- adaptive sampling (in-course learning),
- directed visualisation,
- reliance on approximation (reject optimization),
- divided labour (people, machines),
- divide and conquer (consolidate later),
- exploit context.

The image of statisticians requires some profound changes. First and foremost, our research and advisory activities need to be intensified. We should also make a better use of human resources, strengthen our international connections and work on developing them. Finally, the foundations of statistics should be rebuilt in such a way that computer science is incorporated as an important part statistics and statistical computing may prove very useful here. Thus, we need to actively add sets of mass data.

3. Social aspect of statistics

The present stage of social development is characterized by a rapid growth of information technologies which are included by statistical computing. The economic growth depends to a large extent on information management and the ability to use the achievements of modern technologies. Statistics plays a fundamental role in the processes of collection, analysis, interpretation and sharing of information. However, this fact is fully approved neither by the government nor by the economic circles. The importance of statistics as a field of study and scientific research is also underestimated due to both insufficient popularity among students and limited career prospects available for graduates.

In his article published in the International Statistical Review in 2007 professor Stephen Stigler of the University of Chicago presents arguments which support his claim about the influence of statistics on the prosperity of nations The course of the history of the last 400 years shows that the best developed countries were at the same time the countries which had the best statistical systems. This statement refers not only to Britain but also to many other countries: France, Germany, the Netherlands, Belgium and, later, to particular states of the United States and India. Whether good statistics promotes successful development of a nation, or a favorable growth of a country stimulates the development of statistics is a debatable issue. When we analyze the development of countries from a historic perspective we can observe advances

in statistics as the result of economic achievements of a given country. On the other hand, it is also undisputable that the development of statistics can play an important role in creating the country's economic success. Yet, we need to remember that according to Adam Smith free market, natural resources and human capital are the main factors of economic development (for details see Stigler, 2007).

Numerous examples can support the statement that the use of statistics can bring considerable profits while the outlays are relatively low. The application of statistical methods for planning experiments in many areas of the processing industry in the United States in the 60-tiescan be quoted here. A well functioning system of national statistics can be used as a factor influencing the increase in social efficiency. Moreover, it can promote a substantial reduction in manufacturing loss, increase in service range, stability of economic policy and a better evaluation of future needs.

As a result of a long-lasting process of historic development, statistics (e.g. in Canada) has managed to develop a set of tools which can be used effectively to solve numerous problems connected with streamlining the functioning of a company. Unfortunately, not many of these tools are used in practice. At the company level the main aim of statistics is to support the management in execution of tasks carried out at three levels: strategic, managerial and operational. At the strategic level the most important role is played by statistical thinking which consists in the ability to associate different phenomena, taking decisions based on available information, understanding the notion of variability and, finally, the ability to act in an organized and systematic way.

Statistics and statistical computing, which is nowadays present in almost every sphere of human activity, would not be able to function well without the application of statistical methods. At the same time the knowledge of statistics remains almost totally beyond the scope of interest of a vast part of population or is perceived as a highly specialist field of mathematics accessible only to a narrow group of scientists. On the other hand, media constantly provide information on different aspects of human activity: results of economic activity, current situation on the stock exchange or safety of road and air traffic. This information is not always fully comprehended by the general public. Therefore, a need arises to rethink the role of statistics in the areas of education, business, public administration as well as in society. The task seems especially challenging now that societies head for globalization.

"Statistics is a book by which we can climb one rung in the ladd*er* from data to information" (Rao, 1997).

References

Bryce, G.R., Gould, R., Notz W.I., Peck, R.L. (2000). *Curriculum Guidelines for Bachelor of Science Degrees in Statistical Science.* The ASA Undergraduate Statistics Education Initiative Website: http://www.amstat.org/meetings/jsm/2000/usei/BS-curriculum.PDF.

Bessant, K.C., MacPherson, E.D. (2002). Thoughts on the Origins, Concepts, and Pedagogy of Statistics as a "Separate Discipline". *The American Statistician*, *56* (1), 22–28.

Domański, Cz. (2011). Statystyka nauką dla wszystkich. In: *Statystyka w Służbie Publicznej. Wyzwania XXI w*. Kraków: Statistical Office in Kraków.

Givens, G.H., Hoeting, J.A. (2005). *Computational Statistics. Wiley Series in Probability and Statistics*. New York: Wiley-Interscience.

Lauro, C. (1996). Computational statistics or statistical computing, is that the question?, *Computational Statistics & Data Analysis*, *23* (1), 191–193.

Nolan, D., Temple, L.D. (2010). Computing in the Statistics Curricula. *The American Statistician*, *64* (2), 97–107.

Rao, C.R. (1997). *Statistics and Truth: Putting Chance to Work* (2nd Edition). Singapore: World Scientific Publication.

Rao, C.R. (1994). *Statystyka i prawda*. Warszawa: Wydawnictwo Naukowe PWN.

Stefanowicz, B. (2001). Edukacja statystyczna. *Kwartalnik Statystyczny*, *III* (1), 2–5.

Stigler, S.M. (2007). Statistics and the Wealth of Nations. *International Statistical Review*, *73* (2), 223–226.

Szupiluk, R. (2013). *Dekompozycje wielowymiarowe w agregacji predykcyjnych modeli datamining*. Warszawa: The Publishing House of the Warsaw School of Economics.