Rethinking Statistics with Big Data: learning from George Box

By
Daniel Peña
Department of Statistics and Institute UC3M-BS of Financial Big Data
Universidad Carlos III de Madrid
daniel.pena@uc3m.es

Abstract

In this comment on the article by Prof. M. F. Ramalhoto:" In Memoriam of George Box and a View of Future Directions", I argue that the life and contributions of George Box offer useful insights about how the quality and quantity of available data has led to some important changes in Statistics. I believe that the Big Data revolution will have a strong impact on the evolution of applied statistics and data analysis, fields in which George Box was a master.

George Box was a great statistician and I learned a lot about Science and Statistics from his brilliant conversations, his personality and his writings. In fact, he has been one of the most interesting personalities I have ever met. For this reason it is a great pleasure to contribute to his memory and to comment on the interesting article by Prof. M. F. Ramalhoto:" In Memoriam of George Box and a View of Future Directions". Prof. Ramalhoto wonders if the Big Data revolution will transform Statistics and will have a strong effect on the quality movement. I think it will. George Box always insisted in the need to combine statistical theory and data analyses to solve scientific problems. He strongly believed, based on his own experience, that data was the key ingredient to determine which models we can imagine and fit and used the revolutionary contributions of Fisher (Box, 1976) to illustrate how facts, (data) can lead theory (models) and the needed interaction between both worlds. He said in this paper: " A proper balance of theory and practice is needed and, most important, statisticians must learn how to be good scientists; a talent which has to be acquired by experience and example".

This approach to Statistics explains the originality, inventiveness, and importance of George Box's contributions, that I have reviewed elsewhere (see Peña, 2001, 2002). In this note I will briefly analyze why some of these main advances were driven by the available data to solve the problems he faced. Then, from this analysis, we can foresee some future changes in Statistics that will be driven by the opportunities that Big Data will provide to data scientists.

The Design of Experiments was created by Fisher when working at Rothamsted Agricultural Experimental Station. Fisher had to wait several months (or a year) to know the yield of a plot in which he had changed the experimental conditions and, therefore, it was crucial for him to obtain measurements with the maximum amount of information about the experimental process. On the other hand, George Box was working at Imperial Chemical Industries, where he received continuously data on the performance of a chemical process. Thus, it is not surprising that he approached experimental design in a different way than Fisher. His proposal, Evolutionary Operation (EVOP), was based on the idea that any industrial process, in addition to providing some product, produce also data, from which we can obtain information about how to run it in a more efficient and useful way. Box proposed running simple factorial experiments based on small modifications on the key variables of the process. Analysing the effect of these variables on the output may lead to new knowledge for improving the performance of the process. These ideas are today as valuable as 50 years ago, but now we can carry out all the EVOP procedure in an automatic way, measuring simultaneously hundreds of ouput variables and identifying the interesting changes for improving the process. In fact, now we can handle a large number of variables with decreasing cost, because many of the relevant measurements can be carried out in an automatic way. However, we need to develop more powerful procedures to analyse the effect on the output (which usually will be a multivariate dynamic variable) of a very large number of variables that are observed in continuous time. These effects may be non linear and we need to include the dynamics so that functional data analysis seems like a promising tool for the new EVOP techniques.

Fisher was interested in taking advantage of every possible data point, because according to his experience, and the state of technology at his time, every measure obtained was a costly decision. Thus, efficiency, which implies obtaining the maximum amount of information, was crucial in his statistical outlook. For instance, this was the main reason to recommend maximum likelihood estimation. When more data were available, as it was the case in many of the problems considered by George Box, in addition to worrying about efficiency we need also to consider robustness: Outliers due to experimental error or heterogeneity in the process can destroy the optimality of the estimate and lead to very poor estimation and forecasting results if they are not detected or allowed for. For that reason Box invented and developed the idea of robustness, jointly with John Tukey, another giant of the theory-practice interaction loop. Some heterogeneity in the data was common in the data sets at the end of last century but in the large data sets we are dealing with nowadays heterogeneity is the norm and often the main problem. In fact, the most common situation in a data set will be a mixture of different populations or clusters. Identifying these clusters must be the first step in a statistical analysis. This new approach to descriptive statistics will lead to important changes in the way Statistics is thought and applied in practice.

George Box also played a key role in the integration of the classical and Bayesian schools of thought in Statistical Inference (Box, 1980). In this area he was a real pioneering because the blended of classical and Bayesian ideas are at the root of many machine learning procedures for Big Data. Note that the availability of different sets of large data of unequal quality and precision requires Bayesian ideas to integrated these heterogeneous sources of information.

Time series data in the first half of the XX-th century was mostly of rainfall, temperature and other variables linked to the weather. It is not surprising that spectral analysis was the main tool for time series analysis. Holt-Winters in their analysis of industrial and business time series introduced exponential smoothing and open the door to time based approaches (see Box, 1991 for a nice explanation of exponential smoothing). In the 60's Box and Jenkins proposed practical procedures for modelling and forecasting time series using ARIMA models, which are now used in all fields of knowledge. Their approach, or ARIMA modelling, is learnt today by all students of Statistics, Economics, Engineering or Social Sciences. Very few scientific developments of the last 50 years have had such a big impact over the whole scientific community. To give an idea of his influence in economics, today all major central Banks and Statistical agencies are analysing economic data and forecasting them by using the methods he developed with his co-authors.

Large time series are often heterogeneous and non linear and linear models, as ARIMA, are not always appropriate for time series with observations taken with large frequency, as for instance with stock prices. Also, multiple time series are now more common and procedures to reduce the dimension of a vector of time series have become very important, a field in which, again, George Box was a pioneering (Peña and Box,1987).

In the last part of his life George Box produced a major contribution to the Statistical Quality field. He had a key role in showing the importance of statistical thinking in improving quality and productivity, and he developed many statistical tools and procedures widely used in industry. One of his important contributions is the idea to modify the standard statistical criterion used in statistics, mean squared error, and introduce some smoothing through the cost function so that the resulting solution is close to optimal and much more robust and useful to use in practice (see Box and Luceño, 1997). This smoothing idea has been applied with great success to statistical learning (see Hastie, Tibshirani and Friedman, 2009).

Prof . Ramalhoto discusses in her paper the importance of the combination of Stochastics, Science and Engineering (SSE) and some reasons for a possible paradigm shift in the quality movement. In my opinion the key ingredient of the next paradigm shift will be the Big Data revolution, which will bring new opportunities for innovation, creative thinking and scientific discovery in data science and applied statistics. Data will always be the raw material for statistics, as the contributions of George Box have shown.

References

Box, G. E. P. (1976) “Science and Statistics” *Journal of the American Statistical Association*, 71, 356, 791-799

Box, G. E. P. (1980) "Sampling and Bayes' inference in scientific modeling and robustness" *J. Roy. Stat. Soc., Series A,* 143, 4, 383-430.

Box, G. E. P. (1991) "Understanding exponential smoothing: A simple way to forecast Sales and Inventory" *Quality Engineering, 3, 4,* 561-566.

Box, G. E. P. and Luceño, A. (1997). *Statistical Control By Monitoring and Feedback Adjustment*. Wiley.

Peña D. and Box, G.E. P. (1986) “Identifying a simplifying structure in time series”. *Journal of the American Statistical Association*. 82, 399, 836-843. 1987.

Peña, D. (2002) “The Major Contribution of Professor George E. P. Box to Applied Statistics”, *Chemometrics and Intelligent Laboratory Systems,* 63, 5-6.

Peña, D. (2001) “An Interview with George E. P. Box”, *International Journal of Forecasting*, 17, 1-9.