Journal of
Business & Financial Affairs

Das, J Bus & Fin Aff 2012, 1:4
DOI: 10.4172/2167-0234.1000e119

Editorial Open Access

Statistical Thinking: From "Small data" to "Big data"

Dr. Kiranmoy Das*

Department of Statistics, Temple University, USA

Statistical thinking has dramatically changed in recent years when the advanced technology provides "Big" datasets and the challenge is to handle such ultrahigh dimensional data. Since Statistics is not a basic science but an applied science in true sense, this change of thoughts is mainly due to the advancement of basic sciences and data collection procedures.

Statistical science is not yet in its adult stage. The fundamental works in the area were done only in last century by Fisher, Karl Pearson, Neyman Pearson, C.R. Rao and some other notable persons. Prof. R.A. Fisher who might be called the "father of the modern statistical thinking" was indeed a geneticist. His major works were motivated strongly by real applications from various scientific areas like Genetics, Biology, Agriculture, Physics etc.

From the middle of the last century, theoretical statistics was developed by the statisticians who were trained in Mathematics. Mathematical justifications and theories were indispensable for universal acceptance of the basic statistical theories. Although the "Large sample" Statistical theory is not new but the motivation for such theory was definitely not "Big datasets". Most of the applied works in last century were based on small sample idea and "Asymptotic Theory" was just a fancy mathematical manipulation for the then mathematical statisticians. Historically, that was pre-computer age.

Last 20 years of the 20th century must be called "golden period" for statistical science. Statistical computations played a major role in scientific researches, not limited to basic sciences but also in Finance, Marketing, Business and Management. Dimension reduction was the major focus for theoretical statistical research at that time. Computer-based methods like Bootstrap [1], Lasso [2], MCMC [3] become really popular and appealing to the applied scientists. It must be noted that the traditional "Asymptotic Theory" become more relevant and useful under such environment.

MCMC approach which is mainly Gibbs sampler and Metropolis-Hastings algorithm created a new domain in Bayesian Statistics. Bayesian philosophy was quite acceptable in Statistics from the very beginning but its application was limited to a few classical simple models. That was mainly because of the computational difficulty of Bayesian approach for non-standard complex models. MCMC approach just unlocked this door and complex Bayesian models were successfully used for real problems. In recent years, Bayesian models are used in Spatial Statistics to predict climate change for certain environment, in Genetics to locate genes controlling some important traits and diseases [4,5] and in many other interesting disciplines.

Successful completion of "Human Genome Project" (2003) and its derivative "Hapmap Project" (2005) made a revolution in genetic research in the beginning of the 21st century. Millions of Single Nucleotide Polymorphisms (SNPs) were mapped and the challenge is to handle such ultrahigh dimensional data in the popular Genome-Wide Association Studies (GWAS). Dimension reduction, variable selection, multiple hypotheses testing etc. many such new statistical concepts are used successfully in GWAS and complex datasets also motivate the researchers to develop advanced methodology.

This must be the beginning of the adult stage of Statistical science. Big data problem is a common issue in many disciplines and Statistical techniques are extremely useful for the analysis of such datasets. Statistics is shaking hands with Computer science for developing theoretically sound algorithms to handle big datasets. Interdisciplinary research is encouraged and acknowledged in the scientific community today and statistical thinking is very important for such collaborative works. In one direction, statistical tools will be used for solving real problems which will make the subject meaningful to everyone. In the reverse direction, complex datasets will encourage theoretical researchers to develop sound statistical theories which can be applied immediately by people from other disciplines.

References

1. Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. Ann Stat 7: 1-26.
2. Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. J R Statist Soc B 58: 267-288.
3. Geman S, Geman D (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans Pattern Anal Mach Intell 6:721-741.
4. Das K, Li J, Wang Z, Tong C, Fu G, et al. (2011) A dynamic model for genome-wide association studies. Hum Genet 129: 629-639.
5. Das K, Li J, Fu G, Wang Z, Wu R (2011) Genome-wide association studies for bivariate sparse longitudinal data. Hum Hered 72: 110-120.

***Corresponding author:** Dr. Kiranmoy Das, Department of Statistics, Temple University, USA, E-mail: kiranmoy.das@temple.edu

Received October 10, 2012; **Accepted** October 11, 2012; **Published** October 20, 2012

Citation: Das K (2012) Statistical Thinking: From "Small data" to "Big data". J Bus & Fin Aff 1:e119. doi:10.4172/2167-0234.1000e119