K C Chakrabarty: Uses and misuses of statistics

Address by Dr K C Chakrabarty, Deputy Governor of the Reserve Bank of India, at the DST Centre for Interdisciplinary Mathematical Sciences, Faculty of Science, Banaras Hindu University, as part of the 150th Birth Anniversary Celebrations of Mahanama Pandit Madan Mohan Malviya, Varanasi, 20 March 2012.

*   *   *

Assistance provided by Shri Abhiman Das in preparation of this address is gratefully acknowledged.

1. Prof. Umesh Singh, Coordinator, DST Centre for Interdisciplinary Mathematical Sciences, Prof Sengupta, Dean Faculty of Science, Prof Joshi, other distinguished members of faculty of the University, and above all, dear students. I thank you all for inviting me to be in your midst during the 150th birth anniversary celebrations of Mahamana Pandit Madan Mohan Malviya. It is a great honour and privilege for me.

Pandit Madan Mohan Malviya

2. You have provided an opportunity to me to pay my tribute to the "Mahamana" by delivering a lecture as part of the celebrations of his 150th birth anniversary. He was one of the greatest personalities that this nation has ever produced. In many senses, he was what most of the top class institutions today vie to produce. He was, at the same time a great patriot, an eminent educationist, a teacher of teachers, a silver-tongued transcendental orator, an ancient as well as a modern leader, a reluctant but an eminent lawyer, a social reformer, a great human being, a torch-bearer of the downtrodden, and above all, a great nation builder. He lived his life for us and for the future generations. Each and every alumnus of BHU will remain ever grateful to him for having created this university, a capital of all discipline.

My association with BHU

3. Of course, some of you know of my association with this esteemed University and the city of Varanasi. I stayed in this city from my early childhood, did my schooling here and then joined the BHU for my graduation and post graduation. Then, I became a lecturer here while I pursued my Ph.D. I taught here for about 5 years and regard that time as one of the best periods of my life. Though my career progression in the banking system including joining the RBI was not as a statistician, I must confess that the analytical ability and articulation skill based on sound knowledge of statistics, which I developed during my academic career both as a student and teacher of Statistics at BHU, greatly contributed to my progress and success in banking career. I am grateful and thankful to all my teachers and colleagues in the department for what I am today.

4. Today, I have chosen to speak on a theme which is directly relevant to all of you and that is "Uses and Misuses of Statistics". Let me begin with a quote:

> *The safety of science depends on the existence of men who care more for the justice of their methods than for the value of any results obtained by using them.*[1]

5. Statistics is a method of learning from experience and decision making under uncertainty. That is why it is often called the study of the laws of chance. Chance is inherent in all natural phenomena, and the only way of understanding nature and making predictions

[1] Cohen, Morris R., on "Scientific Method," in The Encyclopaedia of the Social Sciences.

is to study the laws of chance and formulate appropriate rules of action. Chance may appear as an obstructor and an irritant in our daily life but chance can also create. Through statistics, we have now learnt to put chance to work for the benefit of mankind. C. R. Rao, in the preface of his famous book "Statistics and Truth" thus said "*all knowledge is, in the final analysis, history. All sciences are, in the abstract, mathematics and all methods of acquiring knowledge are essentially statistics*."

6. The importance of studying the nature of uncertainty was realised centuries ago. In particular, the introduction of statistical ideas in physics began with the need to deal with errors in astronomical measurements. It was Galileo (1564–1642) who first realized that repeated measurements under identical conditions do vary. Indeed he emphasized that *"measure again and again to find out the difference and difference of the difference"*. About 200 years later, Gauss (1777–1855) studied the probability laws of errors in measurements, which was finally formulated in what we today call the normal distribution.

7. The centrality of uncertainties in pursuit of knowledge did not find enough support from the then scientific community. It is said that even Einstein was initially of the view that theories whose objects are connected by laws should be based on facts, not probabilities. He once famously said, *"……God does not play dice with the universe….."* But surprisingly, he accepted the chance behaviour of molecules suggested by S. N. Bose, which resulted in the Bose-Einstein theory. The inherent nature of uncertainties was initially less understood. Although there are uncertainties at the individual level, very few realised a certain amount of stability in the average of a mass of individuals. This is what we call as "order in disorder", or more technically "statistical regularity", which paved the way for formulation of law of large numbers in statistics.

Discipline of statistics: ancient roots

8. Statistics as subject of collection and use of various kinds of statistics can be traced much earlier in India. Kautilya's Arthasastra (321–296B.C.), one of the greatest treatise of economics, indicates a system of census and data collection relating to agriculture, population and other economic activities, covering villages and towns. In addition, the concept of cross-checking and validation by independent agents was very much part of the data collection system. Evidences in the writings of Huen Tsang of seventh to early eighth century and Ain-i-Akbari by Abul Fazal show enough empirical basis for handling administration and politics. Essentially, the basic statistical system remains with the government. Modern statistical information system was developed by the British with the objective of collecting taxes and does not meet with the developmental and administrative needs of the society to a large extent. The commercial and business information system were developed primarily for accountants and does not serve the purpose of management. Hence, the statistical system has to reorient to meet the developmental, administrative and management needs of the society. Practitioners of statistics have to play an important role in achieving this objective.

Statistics pervades our daily lives

9. We all live with statistics. Starting from the morning newspapers to the evening TV reports, we are surrounded by various statistical figures. We often come across data about out *economy: India's population rose to 1.21 billion people over the last 10 years, an increase by 181 million, according to the Census 2011; India's GDP increased by 8.5% in 2010–11; Inflation was about 7% in February this year. Deposits of scheduled commercial banks was up by 14.6% in February, 2012 over a year ago.* How are these numbers generated? All these figures are themselves part of statistical measurement. There are elaborate statistical systems to collect detailed data and compile them to generate the aggregate numbers.

10. Statistics provides a better understanding and exact description of a phenomenon of nature. Through an appropriate framework of data collection and well-organised planning of a scientific inquiry in any field of life, it helps in presenting complex information in a suitable tabular, diagrammatic and graphic form for an easy and clear comprehension, and helps in drawing valid inference about the population parameters from the sample data. If a problem can be properly formulated and measurement data can be generated, whether it arises in physical, biological, social sciences or any other discipline, statistical tools can be designed to provide a scientific solution. Thus, it is widely recognized that the proper use of statistics is a key element of scientific enquiry. At this stage let me emphasize that quality and integrity of data is the most important element in the success and utility of statistics.

Uses of statistics

11. Let me give some examples of its uses. It plays a critical role in agriculture in deciding the plant varieties, combination of fertilizer, pesticides, densities, soil qualities, and growth of output. India's statistics exploration started with anthropological experiment. Without the use of statistics, business and economics cannot make proper planning and policy. Performance measurement in education, including IQ, is a statistical construct. Use of statistics in psychological behaviour is known as psychometry. Examination of climate change and environmental studies are effectively statistical data analysis of weather and environment. One of the simple applications of statistics is to estimate the number of fish in a pond, using random sampling and suitable statistical (*hypergeometric*) distribution. Does the data support genetic theories of inherited characteristics? What is the pattern of rainfall? What are the important risk factors for heart disease? Any marketing strategy is based on some statistical sample study. Is education attainment related to income and health? The Duckworth-Lewis method implemented in limited over cricket is a statistical method. I can keep on giving examples.

Misuses of statistics

12. But since its beginning, statistics has also been misused and there has been considerable debate about how to understand or respond to the misuse of statistics. After all, who hasn't heard of the famous phrase: *"There are three kinds of Lies: Lies, Damned Lies and Statistics*", which is variously attributed to Benjamin Disraeli, Alfred Marshall, Mark Twain and many others. To understand what is meant by "misusing statistics," it is important to describe the role of statistics in the scientific method and relate the concept of "misuse" to other ethical concepts, such as "misconduct" or "incompetence" or "negligence." Some misuses of statistics can be considered misconduct, although most misuses should be viewed as negligence or deficits of competence. While on the ethical dimension, I shall elaborate further, I shall deal with the other dimensions of "negligence" and "incompetence" as part of information illiteracy slightly later.

13. How statistics has been used and what could be its consequence of misuse is well summarised in the Preamble to the Ethical Guidelines for Statistical Practice[2]. It says the *professional performance of statistical analyses is essential to many aspects of society. The use of statistics in medical diagnoses and biomedical research may affect whether individuals live or die, whether their health is protected or jeopardized, and whether medical science advances or gets sidetracked. Life, death, and health, as well as efficiency, may be at stake in statistical analyses of occupational, environmental, or transportation safety. Early detection and control of new or recurrent infectious diseases depend on sound*

[2] American Statistical Association. Ethical Guidelines for Statistical Practice. Alexandria, VA: American Statistical Association, 1999. Available at http:// amstat.org/profession/ ethicalstatistics.html.

epidemiological statistics. Mental and social health may be at stake in psychological and sociological applications of statistical analysis. Effective functioning of the economy depends on the availability of reliable, timely, and properly interpreted economic data. The profitability of individual firms depends in part on their quality control and their market research, both of which should rely on statistical methods. Agricultural productivity benefits greatly from statistically sound applications to research and output reporting. Governmental policy decisions regarding public health, criminal justice, social equity, education, the environment, and other matters depend in part on sound statistics. Scientific and engineering research in all disciplines requires the careful design and analysis of experiments and observations. To the extent that uncertainty and measurement error are involved – as they are in most research – research design, data quality management, analysis, and interpretation are all crucially dependent on statistical concepts and methods. Even in theory, much of science and engineering involves natural variability. Variability, whether great or small, must be carefully examined both for random error and for possible researcher bias or wishful thinking. . . . Because society depends on sound statistical practice, all practitioners of statistics, whatever their training and occupation, have social obligations to perform their work in a professional, competent, and ethical manner.

14. The results of statistical investigations are usually stated in numerical form and are, therefore, in the public mind, assigned a degree of definiteness usually associated with mathematical technique. The careful investigator, however, is constantly aware of the fact that the preciseness of his numerical result varies directly with the degree of care used in selecting, from the larger universe, the sample upon which his study is based. The numerical conclusions derived from a study of the sample are held to be characteristic and representative of the universe. A very common error in statistical investigation is the selection of a sample which is not an accurate cross-section of the larger universe but merely a particular, unique segment. Conclusions drawn from the biased sample will not, of course, accurately reflect the larger universe. The misuse occurs when such conclusions are held to be representative of the universe by those who either deliberately or un-consciously overlook the sampling bias.

15. An article about cats appeared in *The New York Times*' on August 22, 1989. It stated, "*The experts have also developed startling evidence of the cat's renowned ability to survive, this time in the particular setting of New York City, where cats are prone at this time of year to fall from open windows in tall buildings. Researchers call the phenomenon feline high-rise syndrome*." Statistics were like this: from June 4 through November 4, 1984, 132 such victims were admitted to the Animal Medical Centre and most of the cats landed on concrete and most survived. From the data on the distance of the fall for 129 of the 132 cats, it was observed that the falls ranged from 2 to 32 stories. Only one of 22 cats that plunged from above 7 stories died, and there was only one fracture among the 13 that fell more than 9 stories. But how a cat will survive of a fall from a great height defying gravitation? Such description generally does not push one to scrutinize the statements till it was understood that majority of the cat owners do not report these incidents to any medical centre and believe that other people probably don't report their cats' deaths, either. Therefore, the error seemed so obvious that sample was not representative and there was data reporting problems.

16. Let me cite another example on the wrong notion of conditional probabilities, which appear in *Statistical Science* 2005, a type of intentional and unintentional misinterpretation of statistical information. It says when discussing fatalities on highways on page 78: "Four times more fatalities occur on the highways at 7 p.m. than at 7 a.m." This of course does not imply, as some newspaper had suggested, that it is more dangerous to drive in the evening than in the morning. Recast in the language of conditional probabilities, that *P(*accident | 7 p.m.*)* should not be confused with *P(*7 p.m. | accident*)*. Unfortunately, it was. One more example, a study of the deaths arising out of road accidents revealed that 98 percent deaths occurs while driving on the left side of the road, whereas, only 2 percent death occurs in case of

people travelling in the middle of the road. Hence, it was wrongly inferred that it was safer to travel in the middle than on the side of the road.

17. Another area of concern has been the misuse of statistics coming from spurious correlation and attributing such relation to type of cause and effect phenomenon. In most cases a common variable, most importantly "time" works as a link. Let us take another example. Number of people watching TV serials and number of buffaloes will register a high correlation. If one tries to link them and finds a causal relationship is simply absurd. It is easy to get carried way. One of the prime misuses of statistics is finding a strong relationship between two variables when actually such correlation is spurious. However, exploration of such behaviour has also led to path breaking developments in econometrics. For example, most of the macroeconomic series have strong time trend and thus have strong correlation. Prima facie one may suspect the existence of spurious correlation. However, further exploration in such relationships led the foundation of the concept of cointegration. If two variables have a common trend, then there could be a possibility of a long-run equilibrium relationship between them. This statistical finding opened up a new dimension of modern economics and CWJ Granger was awarded Nobel prize in economics for this in 2003.

18. Let me now give you an example from the banking world. The concept of inflation is often misunderstood by people. You would often hear this refrain from many persons that despite RBI claiming that inflation has come down in the recent past, the prices have not come down. So, what is the truth? Well, the truth is simply that inflation is indicated as a percent change, so even when this percent is declining, all it means is that the prices are rising at a slower rate, but they are still rising! The prices would come down only when the inflation becomes negative.

19. In conclusion, we must ask ourselves as to why a quotation like "Lies, Damned Lies, and Statistics" as mentioned by me earlier is made about the subject of statistics. In my opinion, it is because the society is illiterate, i.e. it is information illiterate. Information literacy, which most people misunderstand, is the third generation of literacy. The first generation of literacy is when you know how to read and write. The second generation of literacy is when you are computer literate, but it is not enough to be just first or second generation literate. When transiting to a knowledge society, it is critical for all of us to be information literate or be third generation literate. And, one of the purposes of statistics is to bring about information literacy in the society. If that does not happen then statistics can be used to prove or disprove anything and it is the subject of statistics which would bear the burden of ridicule. So, it is up to all of us, faculty, students and practitioners alike, to redirect our efforts towards spreading information literacy in the society. Institutions associated with teaching and training of statistics have a more important role to play in that direction, and I am confident that students, teachers and practitioners of statistics associated with Department of Mathematical Sciences at BHU will take greater responsibility in spreading information literacy in the society. And this will be your greatest service to the memory of "Mahamana" and to the discipline of Statistics. I wish you good luck in this effort.

Thank you.