Shifting Goals And Mounting Challenges For Statistical Methodology

Pranab K. Sen

University of North Carolina, Chapel Hill

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation

INVITED ARTICLES
Shifting Goals And Mounting Challenges For Statistical Methodolgy

Pranab K. Sen
Departments of Biostatistics and Statistics,
University of North Carolina, Chapel Hill, NC
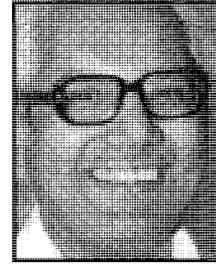
Modern interdisciplinary research in statistical science encompasses a wide field: agriculture, biology, biomedical sciences along with bioinformatics, clinical sciences, education, environmental and public health disciplines, genomic science, industry, molecular genetics, socio-behavior, socio-economics, toxicology, and a variety of other disciplines. Statistical science has historically had mathematical perspectives dominating theoretical and methodological developments. Yet, the advent of modern information technology has opened the doors for highly computation intensive statistical tools (i.e., software), wherein mathematical aspects are often de-emphasized. Knowledge discovery and data mining (KDDM) is now becoming a dominating force, with bioinformatics as a notable example. In view of this apparent discordance between mathematical (frequentist as well as Bayesian) and computational approaches to statistical resolutions, and a genuine need to formulate training as well as research curricula to meet growing demands, a critical appraisal of statistical innovations is made with due respect to its mathematical heritage, as well as scope of application. Some of the challenging statistical tasks are illustrated.

Keywords: Bioinformatics, Biostatistics, Clinical trials, Computational sequence analysis, Data mining, Dosimetry, KDDM, Pharmacogenomics, Quality of life

Introduction

Unlike most of the basic sciences, the evolution of statistical science has followed a somewhat different tract. Indeed, the question can be raised if statistics can be regarded as a science discipline. This might be partially due to its genesis and partially to the increased demand for statistical tools in almost all walks of life and science.

Statistics aimed to capture the power of observations mostly from empirical studies (as in clinical as well as other experimental sciences), and yet to formulate a general body of theory and methodology that provides rationality in a scientific and objective way. Striking a balance between mathematical exactness and natural diversity (variation) has always been the basic goal in statistical reasoning. Nevertheless, combining empirical and methodological reasoning has never been an easy task; it has led to a lot of diversity in the current visualization of statistical science as an integrated and interdisciplinary field. At the same time, it has created some distinct division within the discipline (such as mathematical statistics, probability theory, applied probability theory, stochastic processes, applied statistics, experimental statistics, biometry, biostatistics, actuarial statistics, industrial statistics, public health statistics - notably, epidemiology -, environmetrics, nutrition, health administration and health care, and medical statistics, among others). These sectors are often demarcated with conventional frontiers of their own that do not overlap much.

As a result, the pedagogical, research, and practice portfolios in these sub-disciplines are often not so much compatible. Yet, it is undeniable that statistical reasoning is essential in all these disciplines, and hence, there must be a unifying factor in the statistical foundations. Moreover, the genesis of theoretical statistics (not to be mislabeled with the calculus of variations) is in a consortium of such as gambling, demography, genetics, anthropometry, agriculture, and biology as a whole. (Otherwise, statistics would have probably been a subversion of quantum physics or functional analysis!) It is also undeniable that in order to survive as an objective and scientific way of reasoning, statistics, as a discipline, needs the utmost

Pranab Kumar Sen is Cary C. Boshamer Professor, University of North Carolina, Chapel Hill. He is a Fellow of the American Statistical Association, and the Institute of Mathematical Statistics. He was awarded the 1998 Czech Union of Mathematicians and Physicists Commemoration Medal. His research has mostly been in nonparametrics, covering bioassays, psychometry, multivariate analysis, linear models, sequential analysis, survival analysis, toxicology, bioenvironmetrics, and stochastic processes. He has (co-)authored 10 texts and monographs and 540 publications, and (co-)edited 10 scientific volumes. He also cherishes the opportunity to indulge in literary works, and has published some of his poems.

patronage of mathematical logic as well as understanding. This remains pertinent to data analysis in a valid statistical way, and much of the discussion to follow will center on this point.

Is it natural to see (even) elementary school children playing with personal computers, and start data analysis (in the pretense of statistics) without having any idea of the subject matter? Perhaps, the advent of information technology might have changed the interface of statistical science in this way. If so, coping with data analysis may be the future goal of statistical science!

From a pedagogical perspective, (with the tendency of specialization in a of any broad discipline), it becomes imperative to examine undercurrents carefully, and to remodel programs accordingly. This relates to the necessity of examining the structure of high school level training in mathematical sciences with a view to orient the prospective college-bound students in statistical science (of course, with a good amount of planning so as not to create an imbalance in the training of people in this area with suitable job prospects and career development opportunities), to introduce the basics of statistical science at the high school to undergraduate levels, and to promote graduate level training for those who reach that level.

In this respect, the coverage of the subject material as well as students' mathematical levels needs to be planned and reviewed in a careful manner. In view of the enormously large number of college-bound students and the diversity of the subjects (compared to fifty years back), as well as, the differential job prospects (that evolutes quite dynamically), there are good reasons to appraise the conventional high school, undergraduate, and graduate curriculum in statistical science. The need for professional (licentiate) statisticians should be addressed as well, as discussed below. Perhaps, it would be more appropriate to depict the evolution of statistical sciences first, and then, in the light of that, to discuss the pertinent issues in other related items.

The Evolution of Statistical Science

A history of the evolution of statistical science could well be dependent on who wrote it and what was to be highlighted in the compilation. Evolution of probability theory and stochastic processes mostly took place in mathematical and physical sciences. It is clearly identifiable that the abstractions came generally from mathematics, but useful applications came generally from physical and engineering sciences. The classical Brownian motion and Brownian bridge are concrete examples in this respect.

Also, statistical mechanics reflects the interactive nature of classical probability theory and quantum mechanics in physics. The field of operations research and systems analysis has exhibited a high level of interaction between statistical reasoning with industrial engineering and operations management policies. Yet, a very systematic integration of probability theory and stochastic processes with mathematical statistics did not take place until about seventy-five years ago, and the channel is still quite open.

Of course, long before the evolution of modern mathematical statistics started, statistics arose as a study of the description of states, especially their economic and demographic aspects. A match of this origin of statistics can be traced to ancient China, Egypt, and India among other places. For example, *chautha*, one-quarter of a farmer's harvest, was to be given to the royal (or ruler's) depository, as a practice in India and other oriental places, and this could not have been done without some statistical sense.

Aside from such usages more than several thousand years ago, the more systematic tract of the evolution can be traced in the sixteenth century (CE) where several European countries (or kingdoms) developed some statistical reasoning for various administrative purposes. Common examples include actuarial and demographic usages. However, even astronomy was not outside this realm of statistics. For a nice source of this early historical perspective in statistics, refer to Kendall (1978) and Stigler (1986). They provide good historical accounts of the statistical heritage.

It is much less a problem to assess that developments in statistics in anthropometry, agriculture, genetics, education, psychology, and medicine sectors contributed no less generously to modern statistical theory (which is only just about 100 years old). In that way, mathematical statistics emerged as a discipline that incorporated probability theory and stochastic processes to institute mathematical treatment of data in a rigorous and scientific way, allowing variability to manifest as a natural part of the data diversity. In this way, it also inherited certain basic regularity conditions under which such objective mathematical treatises were adoptable for observational or experimental data models. Alas, it did not perceive the advent of modern information technology that could totally wipe out these methodological landmarks!

With the focus on the evolution of statistical methodology, it would be fair to acknowledge the developments in Great Britain along with contemporary as well as earlier developments in the Continental Europe. In the English sector, the scenario was dominated by the stalwarts: Karl Pearson, Ronald A. Fisher, Egon S. Pearson and J. Neyman, who actually moved to England and finally to USA. (Much earlier, the evolution of Bayes theorem initiated a novel line of attack, and even after a couple of centuries, statisticians are struggling to rationalize the Bayesian philosophy in diverse setups.) The K. Pearson Era could roughly be equated to prior to World War I, while the major thrust of Fisher's work was conducted post World War I period (1920 - 1939).

The centenary of K. Pearson's classical goodness of fit test in 2001 prompted a special coverage of *Biometrika* to fathom the history of allied developments during the past 100 years. The (E. S.) Pearson - Neyman Era, overlapping with Fisher, captured the attention mostly during 1936-1950. Even though the Fisherian *likelihoods* and Neyman-Pearson *hypothesis testing* doctrines still reign in the statistical arena, there are threats for revolutionary changes coming from the knowledge discovery and data mining (KDDM) quarter, (and there is a need to watch out).

Although this evolution mostly in England had a clear biometric, anthropometric and genetics influence, in Russia (later Soviet Union, and now back to Russia) with genesis in gambling (St. Petersburg games), systematic developments took place in the area of probability theory and stochastic processes: A. A. Cuprov, A. A. Markov, V. I. Romanovsky, A. Khintchine, and A. N. Kolmogorov are specially noteworthy in this respect. Much of the foundations of modern probability theory and stochastic processes were laid down by the East European mathematicians, while in the second half of the 20th century, the evolution engulfed Western Europe as well the North American Continent.

There were also earlier developments in Western Europe, and in particular, R. von Mises's work on statistics and probability appears to be contemporary to R. A. Fisher's work; for the galaxy of other statisticians, refer to the historical accounts by Stigler (1986) and others. In the post World War II era, statistical science made its' most noticeable strides. In the USA, A. Wald opened the doors for sequential analysis and decision theory, H. Hotelling's earlier contributions to multivariate statistical analysis paved the way for more statistical methodology in psychometry, biometry and other fields. H. Cramer's treatise of mathematical methods of statistics captured the attention from all over, and the arena of statistical methodological research engulfed the USA with the Neyman school in California, and Hotelling, R. C. Bose, S. N. Roy, W. Hoeffding and H. Robbins in North Carolina.

Mostly due to the pioneering leadership of P. C. Mahalanobis and P. V. Sukhatme, India was also on the global statistical map, with the emergence of R. C. Bose, A. Bhattacharya, S. N. Roy, C. R. Rao, and other contemporary stalwarts. In the Australian continent, E. J. G. Pitman emerged with superb statistical ideas that are still inspiring even after many decades. The next phase started in the mid 1960s when epidemiology, environmental science, medical and clinical sciences, and public health disciplines sought much more statistical analyses, leading to the emergence of modern biostatistics. After three more decades, this has become a completely different world: the Information Technology (IT) era: Where do statistical methodology go from here?

In retrospect, the growth and fall of abstract theoretical statistics could be visualized over the past 100 years. In this respect, the early phases are characterized by the foundation of statistical theory and methodology wherein mathematical concepts were developed and extensively appraised. In the later phases, driven by the natural forces of applications, data analysis concepts entered the arena, and in that respect, some fundamental changes in the basic statistical reasoning took place. This radical change is not unique in statistical sciences. In most of the basic sciences there also has been a significant drift from the theoretical to applied works.

Mathematics is the mother of all sciences, philosophy (logic), civilization, and cultural achievements too; mathematical reasoning is used in everyday life and decisions. The evolution of applied mathematics (including biomathematics) stems from the need of mathematics in other areas of scientific and socio-economic endeavors. Statistical science has its genesis partly in this mathematical universe and partly in the immense need for modeling (quantifying) and decision making in a variety of so-called *inexact* disciplines where uncertainty dominates the decision process, and thereby the exactness of mathematical logic stumbles into impasses. As such, statistical science has emerged as a discipline in its own stochastics clouds (over deterministics). It might not be out of the way to mention that this feature of uncertainty or unaccountable outcomes is also shared in economic decisions, although the nature in which that is handled could be quite different.

Mathematicians and theoretical physicists have bondage to mathematical abstractions and sophistications. Applied mathematicians and applied physicists like to use sophisticated mathematical tools to approximate real life problems. Statisticians are, however, somewhat divided by their diversity and professional outlook. Mathematical statisticians and probabilists have much greater affinity to mathematical concepts (e.g., analysis, abstractions, logic), while statisticians and applied probabilists are more akin to applied mathematicians, though committed to stochastics to a greater extent.

Applied (and bio-) statisticians are more data-oriented, and the information technology is changing the outlook across this broad spectrum. That is why there is the need to have an integral view of the statistical science, preserving its diversity and yet understanding its foundation and basic philosophy. Slicing a broader discipline into smaller chips may knowingly or unknowingly affect the integration of the discipline and create impasses for interactive research and training. No wonder, Professor P. C. Mahalanobis, while laying down the foundation of statistical science in India, 70 years ago, proclaimed the logo: *Unity in Diversity.*

Statistics does not aim to distort the prevailing diversity (variation) in a system; rather, it tries to appraise

the picture in a unified way wherein the deterministic and stochastic aspects are appraised in a sound objective manner along with good indications of margins of possible incorrect decisions. The evolution of decision science to cope with uncertainties has its genesis in statistical science (although, the mode of statistical decisions in this respect is somewhat different than in the classical statistical decision theory). The theory of games, as has been developed with due respect to stochastics and importance of decision making, is a classical model for the incorporation of superb statistical reasoning in game theoretic as well as economic decision theory.

Current Trends: Pedagogical Issues

At this juncture of time, in statistical science, the traditional emphasis on agricultural sciences has been degraded in favor of biological, clinical, behavioral, biotechnological sciences, and public health. Though there is a renewed interest in financial statistics (incorporating sophisticated stochastics), the emphasis on traditional econometrics has been partly shifted to environmental, epidemiological and health management sciences (where there is a significant temporal component), and from classical genetics to microbiology and molecular biology, featured by excessively large dimensional data sets. As a result, in statistics curricula, abstract treatises of measure-theory based mathematical statistics, probability theory, and stochastic processes are giving way to more computational-intensive (informatics) and data-oriented (data analysis) teaching cum research programs all over the world. (Computational) Bayesian perspectives are mingling with numerical and simulation approaches to solve some harder theoretical problems. Faced with this complex, there is a profound need to appraise the following issues:

(1) What a teaching/training program in statistical science (at the high school/ undergraduate/graduate levels should we aim at? How should post-graduation job prospects be related to teaching/training of statisticians? How should statistics, biostatistics, and applied statistics teaching/training programs be synchronized in the light of job market as well as academic prospects?

(2) How do we attune the academic research program in statistical science so as to reflect the diversity and enormous scope of applicability (in an interdisciplinary setup as mentioned before)? There is a genuine need for blending mathematical sciences with dominating stochastics that crop up in these fields.

(3) Should statistical science be viewed as a subdivision of mathematical sciences, or computer sciences, or social sciences? As a matter of fact, can any of these broad disciplines engulf statistical science as a sub-discipline?

(4) How do we standardize teaching/training programs in statistical science so as to incorporate more interaction between theoretical and computational statistics? Information technology has opened up a broad avenue of new approaches to teaching of statistics - what are the pros and cons?

(5) Do we emphasize statistical science as an integral teaching/training program by itself? This needs recruitment of prospective students in statistics program, and synchronization of research with such an innovative program. How do we minimize the risk due to choosing unknowingly an inappropriate training program at the undergraduate college level, and how do we reallocate such mismatched entries?

An appraisal of these issues requires an assessment that might well be beyond the scope of this present article. Nevertheless, let me mention some salient points in this respect. What is the main objective of high school education in a country? Is it simply to produce a hallmark of literacy of citizens? With the change from agronomic to technological cum industrial setups in most of the countries, the high school programs are to be viewed from a much broader perspective, such as:

(1) Providing a basic education to growing citizens so that they feel comfortable in the society they live (culturally, socially as well as economically). The enlivenment of cultural and social aspects is generally brought up by the literature and social science subjects in the curriculum. The basic training in literature and mathematics may be the gateway to their passage to higher education.

(2) In the West, at the high school level, a reasonable amount of vocational training programs used to be the facilitating factors for high school graduates to stand on their own with respect to their daily needs in life. For example, typing skills, small mechanics to master household matters, automobile driving training to meet the daily need of commuting, etc.

(3) Sport activities to develop the team spirit and good sportsmanship.

(4) Religion classes to have socio-cultural understandings, and other related matters.

(5) Vocal as well as instrumental music classes, drama and journalism courses to develop arts faculties.

One essential thing in the whole curriculum is the development of social, community and ethical standards and obligations of high school students for the betterment

of the society and community in which they live. Are we still sticking to this interface of our high school level training? In USA, in the melting pot, there are sometimes cracks that raise the question: what should be the primary goal of high school education, and how should be the curriculum developed to optimize this goal? I guess statistics has a lot to contribute towards this materialization.

In any setup, in the high schools, it would be quite evident to perceive the diversity of the students' family background, their level of motivation, diligence, drive, ability to comprehend academic material in diverse setups, as well as, their aims and ambitions in life. A good high-school program should aim to harness this diversity to the advantage of the society and country as a whole. No wonder that some level of specialization is instituted at the high school level. Students stronger in language skills are encouraged to have more advance placements in that direction, while better science and mathematics students are encouraged for higher education in science, engineering or medicine. Yet, in this diversion, it is apparent that not all high school students would be really ready for college level education; even if ideally all high school graduates plan to proceed to the traditional undergraduate education, will it be possible for colleges to absorb this heterogeneity and have a resolution without compromising their standards and main aims and objectives?

As such, alternative vocational training plans for those who choose not to go for academic college programs are also planned at the high school level. A number of years ago, in Japan, for example, only about fifteen percent of the high school graduates used to go for college education, while a majority for some other vocational training. In Europe, the college bound percentage used to be a bit higher, but not as much as in USA. Even there, many of the high school graduates used to get settled in life in their farming or other family occupations, leaving the luxury of college education to a smaller number of talented and academically oriented ones. The value system on which this was based might have changed drastically during the past three decades.

In addition, it remains to appraise the role of statistical science in this career decision-making task. Very few pupils at the high school level think of statistical science as a career objective, and also, in very few high schools, statistics curriculum may be strong enough to influence the pupil to think of such a possibility. Given this bleak picture, how could one expect a major turnover for statistics majors at the undergraduate level, and in turn, at the graduate level? To place proper emphasis on statistical science as a viable career alternative, it may be essential to sow the seeds at the high school level; that requires properly trained and dedicated teachers, and superb curriculum planning on the part of the school boards, state or nationwide. I am not too optimistic about the prospects.

The quality of undergraduate education is determined by the factors (i) quality of college-bound high school graduates, (ii) quality of faculty at the college level, (iii) curriculum of a program, and (iv) general economic conditions permitting the uninterrupted study environment for the students. It may be the impact of information technology that in the high schools, as well, there is a tremendous emphasis on the use of computing and programming skills in their study plan, so that in that process their drive for analytical thinking might be compromised to a certain extent. This feature has been observed to be persistent in USA, and my sense is that it is more or less the case in many other technologically advanced countries (the Far East Countries are different in this respect).

If such a change has taken place and is likely to continue in the foreseeable near future, it may have a profound impact on the training of mathematical sciences at the college level as well. In that way, there is bound to have some impact on statistical methodology as a part of this curriculum. In the high schools in USA, there is an enormous problem in recruiting adequate number of qualified and dedicated teachers who could assume the responsibility to accomplish the teaching and training task without compromising on the quality. This shortage is also linked to emphasis on nonacademic jobs (because of salary differentials and other factors).

The picture may not be too different in the undergraduate colleges, and there is a need to appraise how far existing faculty can promote the right environment for teaching and training in statistical science (that does not de-emphasize the basic role of statistical methodology). In the late 1980s it became quite apparent to college administrators that in many campuses, the undergraduate programs were very narrow or specialized in a sub-domain. There were general recommendations that liberal arts and science programs be brought back at the undergraduate level, and more specialized programs be relegated to the graduate level of training.

This recommendation certainly deserves a lot of commendation from educators from all over the world. This certainly makes the graduates to deal with real life in a better way, and also to choose their profession in a more thoughtful way (instead of being persuaded by the job prospects in specific areas, knowing that the fortunes do not stick to any specific area for a long time). Many campuses in USA are coping with this change in a subtle way, and in many places, statistics has been designated as a required part of the undergraduate training. Although this statistical-literacy is certainly boosting for people in statistical profession, it also raises the question: How much of statistical methodology should be prescribed for undergraduate non-mathematical sciences students? In what way, statistics acts as a binding power for diverse disciplines in this spectrum? How far statistics should be implemented in an

undergraduate mathematical sciences curriculum? For students in experimental sciences major programs, how such statistics curriculum should be blended to suit both methodology and applications? All these issues merit careful appraisals.

It would not be out of the way to refer to interesting volume edited by Gal and Garfield (1997) dealing with the assessment challenge in statistics education, with varied contributions by a number of educators and researchers in statistics. Also, refer to a follow-up article by Garfield and Gel (1999). They outlined nicely some practical implications for college statistics teachers, and stressed on the following assessment challenges:

(1) Assessment of students in computer-assisted environments.

(2) Assessment of statistical literacy.

(3) Assessment of students' understanding 'big ideas'.

(4) Assessment of students' intuitions and reasoning involving probability concepts and processes.

(5) Assessment of outcomes of 'group work'.

(6) Developing models to use in evaluating and comparing curricula.

(7) Using assessment to determine what students understand after they interact with simulation software packages.

The main role of such an assessment is to enhance student learning, albeit in confrontation with the advances in computational challenges that crop up in college level statistics programs. Their assessment certainly serves a good purpose, though I would like it to take a step further in raising these issues not primarily for the college teachers but more for the educators who actually lay down the foundation of statistical curricula in high schools, undergraduate, graduate colleges, as well as, in professional and vocational schools. Teachers are more like the musicians while these planners are more the scripts that the teachers are to play. Although in their assessment, computational skills occupy a focal point, our motivation is to bring the compatibility of computational and analytic skills, when there are many extraneous factors which could have significant impact on the whole scenario.

As indicated earlier, not all high school students are undergraduate college bound, and many of them show up in community colleges and in other vocational training centers. Because of their intention or inability to pursue a more academic program (often, due to family or economic reasons, other than less impressive academic performance

at the high school), any such training program runs on a somewhat different track (in terms of both time and level). Such vocational trainings serve a good purpose for the community and society as a whole (specially, when there is an acute shortage of such people in the profession), and hence, close attention should be paid to such programs.

For many of these vocational training, a curriculum, though specialized in a specific field (e.g., carpentry, electrical installation, builder's job, car mechanics), the growing need of using PCs and laptop PCs also calls for some basic statistical understanding, albeit with less methodological flavor. This might be more akin to data management and data analysis sectors. Therefore, the basic query arises: how much statistics should be in the instructional packages, and to what extent should the instructors be conversant with this material?

Statistics training at the (post-)graduate level is, in turn, dependent on the flow of interested, dedicated and qualified pupils from undergraduate classes as well as migration from other fields. The demand for traditional mathematical statistics post-graduate programs is likely to dwindle in the near future, and, more likely, biostatistics and other applied statistics programs would be expanded. Again, due to the tremendous influence of the current evolution in biotechnology and in information technology, there is a general tendency for incoming graduate school students to seek admission in some of the related programs than in mathematical or statistical sciences. At the same time, with due emphasis on the basic role of statistical data analysis in a broad spectrum of disciplines (including business management, finance, economics, and social sciences), there is the basic dilemma: how much to emphasize on statistical algorithms and packages without probably motivating them from methodologic justifications?

An answer to this query would depend on how we project the growth (or decline) of enrollment in graduate programs in different areas of statistical sciences (e.g., mathematical statistics, biostatistics, applied statistics, experimental statistics, etc.), and how to blend some of these programs to make this transition smoother. Fortunately, in many places, the mathematical statistics programs are undergoing such changes, and are incorporating more applied curriculum to facilitate the blending of methodology with good applications; the loser may be the abstract theory that might ultimately take shelter in hardcore mathematics curriculum. Over the past twenty years we have been observing a changing pattern of graduate students in USA in statistical sciences, most notably, in biostatistics.

As the number of US graduate students in various disciplines dwindled, even the more prestigious campuses have been flooded with international students, many of whom are from China. Most of these incoming students had good training in mathematics in a rather theoretical fashion, and in many cases, in their undergraduate

programs, they had only mathematics, without much emphasis on other subjects, including statistics. In a changing environment in USA, not only they had to adapt to this new academic environment (which certainly have enriched the field in a way), but also they induce a lot of unanticipated changes wherein algorithms and computer skills dominate over statistical insights and interpretations. Although this model may explain how the interface of graduate studies in statistical science in USA is evolving in a different way than anticipated, it also raises the question: how to motivate immigrants from other disciplines to get tuned to the heartbeats of statistical methods while pursuing a graduate program in statistics?

There are a couple of other pedagogical issues that merit some appraisal. First, in most of the professional schools, namely, education, law, nursing, pharmacology, public health, dentistry, and medical and clinical sciences, more and more emphasis is placed on the use of statistical analysis in research as well as operational decision making. Traditionally, in many of these fields, both pupil and teachers were apprehensive of statistical mathematics, though some of that stuff became a part of their curriculum. To make it more effective, in many places, the instructors are chosen from affiliated statistics - biostatistics departments. This is certainly a very healthy sign. Teaching of basic statistics to such professional students require special care (to smooth out the somewhat different attitudes of students and statisticians at large), and this can be done by people having a sound depth in statistics (not just the packages that might be the 'handbook' for such professionals. At the same time, it needs the knowledge as well as genuine interest on the part of the statistics faculty to appreciate the specific field of application, so that things can be presented to the right audience in the right flavor, and with good interpretations and explanations.

Second, with the tremendous increase in demand of statistical analysis in every walk of life and science, a new class of statistical professionals has emerged; they are called licensed statisticians. A few years back, the American Statistical Association chartered this professional group, and provided broad guidelines for their ability to perform the task they would be asked for. Although it is expected that all these professional statisticians have some good statistical training at the graduate level, it is quite possible that many of them might have migrated from other fields. As such, to facilitate more interaction between good statistical packages and their underlying statistical methodology, it might be better to stress the need for adequate familiarity with statistical methodology so as to perform the licentiated job more adequately.

As pedagogical issues are intricately related to the needs of the society and mankind as a whole, the pedagogical value system continues to change progressively over time with the subtle changes in our social, cultural,

political and economic factors. There are indications that such changes are not that subtle any more, and if so, we must prepare ourselves to possible radical changes in some of these aspects.

Not too long ago, among all disciplines, philosophy used to be the trump card, and in basic sciences, physics and chemistry were the royals. With the march of time, the emphasis on logic and philosophy in pedagogical setups has been downloaded in favor of applied sciences and computers. Basic training in mathematics has also gone through radical changes with more emphasis on computers: the PC's have made the slide-rules obsolete! Students in chemistry classes do not need to memorize the 'periodic charts' of atoms and molecules, nor they need to analyze a compound product by exclusive chemical analysis. High-level computers can simulate complex differential equations, and thereby predict anticipated reactions to a reasonable extent. Given these vast computational amenities, analytical thinking is being gradually replaced by simulation and computational skills.

Faced with this radical change, biological, clinical and other applied areas, once thought to the outside the domain of mathematical reasoning, are gaining momentum and catering for more attention from pedagogical as well as career decisions point of view. At the present time, there has been a sustained effort to promote more understanding of biological, clinical, public health, and environmental sciences through basic interdisciplinary research, and (bio)statistics is a binding force to incorporate quantitative reasoning in this broad interdisciplinary field. Biotechnology, in general, and bioinformatics in particular, are reshaping the sphere of human knowledge and endeavor. Yet, there are many challenging tasks ahead. Most imminent areas in this respect include the following questions and emerging arenas:

(1) Information technology: too much information to disseminate?

(2) Bioinformatics (and medical informatics): how large could a biological system be?

(3) Molecular biology and genomics; biotechnology.

(4) Chemometrics and environmetrics.

(5) Environmental health sciences, toxicology.

Most of these novel areas are characterized by very high-dimensional data sets (where there may be too much of information that can be disseminated in a manageable statistical way). Moreover, sometimes, it might be harder even to reformulate such problems in a way that would permit a convenient statistical way in looking for

suitable resolutions. On top of that there could a lack of understanding on the part of a statistician of the biological intricacies, and on the part of scientists in other disciplines of the statistical intricacies. There is a pressing need for applicable statistical methodology for planning, modeling as well as analysis of such scientific endeavors.

Pedagogical aspects are very crucial in this respect. Without recruiting properly trained or knowledgeable persons for this needed statistical task, there could be genuine problems with drawing valid and efficient conclusions. On the other hand, attracting deserving trainees in this disciple needs proper nurturing at the high-school level, so that the prospective college-bound students do not all line-up with medical, law, engineering, or computer science prospects. Statistical methodologic orientation at the undergraduate training level is an essential ingredient in this venture, and I would like to stress it. Before this assessment, however, a few important points require comment.

Career placement prospects cannot be ruled out as a basic factor in teaching and training curriculums in statistical science. Yet this picture is progressively changing over time. After the World War II, in the British Commonwealth Countries, this curriculum used to be quite broad-based with a mixture of applications in various fields and methodology in the main stream, albeit, with some sacrifice of the abstractions that typically germinated from purely probabilistic and measure-theoretic approaches.

On the other hand, in the United States of America, as well as, in continental European countries, mostly statistics teaching programs were housed in mathematics (or mathematical sciences) curriculums, and as a result were substantially more measure-theoretic and mathematically more sophisticated. Applied statistics or experimental statistics programs were housed in engineering schools or in economics and business schools, and were tilted in the respective directions. Biostatistics programs were very few in number, and used to be in either the school of public health or sometimes in the medical schools.

This segregated picture changed gradually over time. More people in applied statistics programs tried to incorporate more statistical methodology in their applied work, and this led to the evolution of *Technometrics, Biometrics* and other journals. The positive aspect of this evolution could be seen from the increasing usage of statistical analysis in medical studies, and later on, many medical journals instituted a convention that without proper statistical support, empirical findings would not be considered adequate for publication.

A more radical change took place about thirty years ago when in the West, health regulatory agencies (e.g., the National Institutes of Health, and Food and Drug Administration in USA) advocated large scale clinical trials where statistical methodology seemed to have a visible role in the decision making process. The drug research groups and pharmaceutical industries were encouraged to have more statistical expertise for better prospects in their marketing and drug-approval ventures. At the present time, drug-research groups may still employ a significant number of statisticians (not all of whom might have the basic training in statistics).

Whither Statistical Reasoning?

In the preceding section, some of the areas where there is a pressing need for statistical reasoning were mentioned, and there are, in that way, some challenges too. I intend on elaborating on these points in more specific contexts, and then append a general discussion on the basic nature of the shifting goals along with these challenging problems.

(I) Genomics and bioinformatics. DNA and genomes have become household words, and there is a global effort to fathom out the mysteries of the intricate network of some 35,000 genes that constitute the human genome. Aided by biotechnological advancements, genomic science has captured the attention of researchers from all walks of life and science. The chemical words (A, C, G, T) that constitute a DNA sequence exhibit qualitative variation, and on top of that there is very large number of sites. There are certain assumptions regarding the DNA activity, and it is not yet clear to what extent statistical methods (and probability theory) can be justified in this context.

The evolutionary interdisciplinary field of bioinformatics covers various aspects of mathematics, statistics, information technology, and of course molecular biology and genetics to a greater extent. However, at the present, it is not precisely known what constitutes the core of bioinformatics. Ewens and Grant (2001) have a nice way of stating the current status:

> We take bioinformatics to mean the emerging field of science growing from the application of mathematics, statistics, and information technology, including computers and the theory surrounding them, to study and analysis of very large biological, and in particular, genetic data sets." The field has been fueled by the increase in DNA data generation leading to massive data sets already generated, and yet to be generated, in particular the data from the human genome project, as well as other genome projects. Bioinformatics does not aim to lay down fundamental mathematical laws that govern biological systems parallel to those laid down in

physics. Such laws, if they exist, are a long way from being determined for biological systems. (Grant, 2001)

At this stage, the main utility of mathematics in this field is in the creation of tools that investigators can use to analyze data. Such tools involve statistical modeling of biological systems, and therefore, there is a genuine need for probability theory, stochastic processes, and statistics in the context of bioinformatics.

Gene scientists cannot scramble fast enough to keep up with the genomics, emerging at a furious pace and in astounding detail. As such, it would be improper to jump on conclusions based on data analysis alone (even under the ubiquitous KDDM umbrella). There is a genuine need to examine the stochastic evolutionary forces underlying large biological systems, in general, and genomics in particular. This venture turns us to look more intimately into the deeper stochastics aspects of macro biological systems, incorporate biological (and genetic) factors, as needed, and only then prescribe a statistical resolution.

This motivated Sen (2001) in the formulation of Biostochastics: Stochastic modeling and statistical analysis of very large biological (including genomic and polygenetic) data sets. Computational sequence analysis (CSA), large genetic data models, and computational biology, in general (Waterman 1995, Lange 1997), have a genuine need of biostochastics for proper methodologic justification. Let us elaborate this point with some specific models.

In the mode of an external analysis, Pinheiro et al. (2000) considered a MANOVA (multivariate analysis of variance) type statistical test for the homogeneity of several groups of people with respect to their DNA sequences, the primary motivation being the impact of HIV (human immunodeficiency virus) on the distribution of the nucleotides. These data models involve purely categorical responses and are immensely of high dimension.

They relate to extensions of the work of Light and Margolin (1971, 1974), because of high-dimensionality, where *Hamming distance* measure has been incorporated to eliminate some conceptual as well as computational difficulties. There are some genuine problems in pursuing a pure likelihood (or some of its variants) approach, and in that sense, the Hamming distance provides a meaningful resolution; more remains to study regarding its desirable or optimal properties.

Similarly, in the mode of an internal analysis, for testing the hypothesis of independence of mutations in a DNA strand, Karnoub et al. (1999) considered a conditional test, and more work is underway in this direction (Sen 2002a). There are many other interesting statistical problems in this area, and some of these are discussed in Ewens and Grant (2001), Waterman (1995) and others. Below, I discuss the salient features of such models in a

general mold.

(II) The PBPK-PBTK Wanderworlds: Pharmaceutical and drug research activities have spurred up in an international phase. From a scientific approach this involves (1) Pharmacodynamics, (2) Pharmacokinetics, (3) Toxicokinetics, (4) Biomechanistics, and (5) PBPK (physiologically based phramacokinetic) models, to deal with the distribution, concentration, absorption, and ingestion of drugs and pharmaceutical elements in the human system (allowing the diversity of health and other conditions). At this time, from a statistical perspective, the vital components are (a) bioequivalence and relative potency studies, (b) dosage-response regression modeling, and (c) assessing genetic undercurrents. There is a tremendous scope for (bio)mathematics and (bio)statistics to work together. Statistical theory and methodology have the key to the basic understanding, but more and more evolutionary work is needed to bring it down to the level of appropriate adoption.

(III) Environmental Pollution and Health Perspectives. Toxicity abounds in nature, environment and in the modern life style. Respirable suspended air particulate matters (RSAPM) constitute the main sources of air pollution, and are regarded as the carrier of toxicity, mostly inhaled through our respiratory system. There are other forms of environmental toxicity and pollutions (such as water and subsoil contamination), and together they work on human as well as other living organisms in a very complex manner.

Because of the latent nature of a large class of toxic substances that are in the environment, the extreme variability of human metabolism as well as their exposure to such toxic materials, yet unknown nature of many carcinogenic activities, and immense difficulties in effective assessment of prevailing toxicity, an exact mathematical dose-response formula is not appropriate here. Even the classical statistical regression analysis is not tenable here (as the basic regularity assumptions pertaining to modeling and analysis are hardly justifiable in such a case, and the sampling design may also be highly nonstandard).

There is a genuine need to incorporate the physiological (as well as molecular biological) mechanism and reactions of such toxic substances in human body (and mind, too), and in the midst of this highly stochastic and complex uptake-intake process, there is a three-phase statistical task for assessment of the aftermaths of environmental pollution, namely, (1) designing (planning) a scientific study that allows reliable measurement of the toxicity levels taking into spatial as well as temporal variation in a reasonable way, (2) formulating suitable models that are design-consistent and yet amenable to further statistical analysis with due biological and environmental

understanding, and (3)analyzing the acquired experimental and observational outcome in a valid and yet simple way.

The sampling techniques for collection of environmental pollution data and associated auxiliary/explanatory variables are generally more complex, and the usual assumption of independence or homogeneity (i.e., spatio-temporal stationarity) may not generally hold, and as a result, much of the appeal of variograms and other standard statistical tools may therefore be lost in this context. Semiparametric formulations have greater appeal than parametric ones, and yet, the very basic semiparametric model may not fit the sampling design as well as the dose-response regressions.

In view of biological undercurrents, biomechanistic models (including the PBTK models) have greater appeal. Nonparametrics should have greater scope too, but may need a comparatively larger data set, a condition that may not always hold in practical applications. On top of that usually collected datasets by different regulatory agencies may conform to an exceedingly large dimension, so that extraction of statistically informative smaller subset of characteristics may often be a challenging task. Refer to Sen (2002b) for a detailed account of such related statistical problems.

The basic idea is to bring in the SARI (structure-activity relationship information) in the picture. The structure refers to the influx of toxicants (along with their bioconcentration factors) and their mode of uptake by human body. Activity refers to the biological, genetic and biochemical reactions that follow the intake of toxins in the body. The relationship is not directly observable and often is obscured by many physiological and behavioral complexities.

In a conventional sense, in a dose response regression, the dose is identifiable and the response is observable. In this context, the dose refers to the influx of toxicants, and there is a lot of variation in their intake by human body. The response refers to the occurrence of a disease or disorder which are more likely to be influenced by the toxicity influx. Nevertheless, the relationship is not so apparent, and the SARI can provide more biological information upon which a plausible regression model can be postulated.

For the SARI there is a profound need to seek biological and environmental impacts, and for statistical modeling and analysis, it is therefore necessary to take into account these factors as effectively as possible. The real challenge is to incorporate the SARI in statistical modeling and analysis. This needs more complex models, more biological and toxicologic feedbacks, imputation from animal studies (dosimetry), and delicate bioassays. Are we expecting statisticians trained in a conventional way to make breakthrough in this domain?

Conclusion

There is an annexation of a new frontier in statistical science from the ongoing evolution of biotechnology and information technology. This has posed some challenging tasks for statistical theory and methodology to cope with planning, modeling and analysis, genuinely needed for drawing scientific conclusions in a statistically sound manner. Most of these highly nonstandard problems may be characterized with the following features:

(1) generally, (very) high-dimensional data models

(2) non-continuous (i.e., binary/categorical or at best, discrete) response variables.

(3) usually, a complex, confounded network of extraneous (and often unobservable) factors

(4) insufficient justifications for adopting simpler theoretical models

(5) likely misspecification, identifiability, incompleteness, and validity problems

(6) scope for statistical reasoning not usually clear.

The wealth of statistical tools and concepts developed during the past six decades in dealing with statistical planning, modeling and analysis of observational and experimental data models, may be of very limited use in this new field. As such, we need to address our curricula, train our teachers, and motivate and organize our training programs to promote better understanding and interaction of statisticians (at large) with other scientists, and create a more congenial environment for the advance of our knowledge.

Although there has been a sustained growth of statistical packages to meet the increasing demand of data analysis, and also promising developments in statistical learning to provide more rationality to KDDM, it remains to see how these diverse tools can also be aged under statistical methods in a broadly interpretable way. I am not pessimistic in this venture, but we need to prepare ourselves better and do our needed homework in this multifaceted task in an adequate way.

References

Anderson, R. J., & Landis, J. R. (1980). CATANOVA for multidimensional contingency tables: nominal-scale response. *Communications In Statistics: Theoretical Methods, 9,* 1191-1206.

Anderson, R. J., & Landis, J. R. (1982). CATANOVA for multidimensional contingency tables: ordinal scale response. *Communications in Statistics: Theoretical Methods, 11,* 257-270.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, A158,* 419-466.

Cox, D. R. (1995). Discussion of Chatfield's paper. *Journal of the Royal Statistical Society, A158,* 455-456.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models for proteins and nucleic acids.* UK: Cambridge Univ. Press.

Ewens, W. J., & Grant, G. R. (2001). *Statistical methods for bioinformatics: An introduction.* NY: Springer.

Fayyad, U. M. et al. (1996). *Advances in knowledge discovery and data mining.* Cambridge, MA: MIT Press.

Friedman, H. P., & Goldberg, J. D. (2000). *Knowledge discovery from data bases and data mining: New paradigms for statistics and data analysis. ASA Biopharma. Repr, 8(2),* 1 -12.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics, 28,* 337-407.

Gal, I., & Garfield, J. (Eds.) (1997). *The assessment challenge in statistics education.* Amsterdam: IOS Press.

Garfield, J., & Gal, I. (1999). Assessment and statistics education: current challenges and directions. *Inter. Statist. Rev., 67,* 1-12.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction.* NY: Springer.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta analysis.* London: Academic Press.

Karnoub, M., Seillier-Moiseiwitsch, F., & Sen, P. K. (1999). A conditional approach to the detection of correlated mutations. *Inst. Math. Statist. Lecture Notes, Monogr. No. 33,* 221 - 235.

Kendall, M. G. (1978). *The history of statistical method. International Encyclopedia of Statistics, 2,* 1093-1101.

Pinheiro, H., Seillier-Moiseiwitsch, F., Sen, P. K., & Eron, J. (2000). Genomic sequence analysis and quasi-multivariate CATANOVA. In *Handbook of Statistics, 18,* 713 - 746.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard Univ. Press.

Sen, P. K. (2001). Excursions in biostochastics: Biometry to biostatistics to bioinformatics. *Lect. Notes, Academia Sinica, Inst. Statist. Sci.* Taipei, Taiwan.

Sen, P. K. (2002a). Computational sequence analysis and genomes: Statistical perspectives and controversies. *Statistics Canada 2001 Conference.* (Y. P. Chaubey, Ed.). UK: World Sc. Press.

Sen, P. K. (2002b, unpublished). Structure-activity relationship information incorporation in health related environmental risk assessment. *Environmetrics.*

Sen, P. K. (in press, 2000). Air pollution: Statistics and environmental health perspectives. *Environmental Issues and Statistical Perspectives.* Oxford University Press.

Waterman, M. S. (1995). *Introduction to computational biology: Maps, sequences, and genomics.* UK: Chapman-Hall.