LASR workshops and emerging methodologies

K.V. Mardia

University of Leeds

1 Introduction

With the 21st century upon us, there has been general stock-taking/soul searching by scientists and humanists. As a result, various edited volumes as well as papers related to this theme have appeared. (See the references.) Indeed, Brockman (2003, p.2) in his edited volume entitled "The New Humanists - Science at the Edge" states that

"The emergence of this new culture is evidence of a great intellectual hunger, a desire for the new and important ideas that drive our times: revolutionary developments in molecular biology, genetic engineering, nanotechnology, artificial intelligence, artificial life, chaos theory, massive parallelism, neural nets, the inflationary universe, fractals, complex adaptive systems, linguistics, superstrings, biodiversity, the human genome, expert systems, punctuated equilibrium, cellular automata, fuzzy logic, virtual reality, cyberspace, and teraflop machines. Among others "

(At least, DNA, double helix, fractal etc. are now in our day to day vocabulary.)

Of course, statistics is not an exception as far as intellectual hunger goes!. Leeds Annual Statistical Research (LASR) workshops are good examples!!

Partly I have been attracted to this topic since, for the last few years, I have been attending some biological conferences, sensing that there is a future for statistics - (at least) in quantitative biology! I attended, a number of years ago, conferences in image analysis when image analysis was becoming a major emerging area. The message has been similar in other conferences where statistics is not a mainstream topic.

In these conferences, one sees that the use of terms like data processes, data analysis, prediction, estimation hypothesis, etc., but the word "statistics" is "innocently" avoided. When suggest to the presenters, that they are really dealing with a statistical problem, one of the following answers comes back:

"Oh yes, statistics is needed but statistics deals only with small data, but here the problem is for very large data". Or

"Oh yes, we will use statistics when we really want to calculate a measure of uncertainty".

In some sense as a profession, we always seem to suffer from delusion. A remark attributed to John Tukey (in Rao & Szekely, 2000) summarizes this well:

"The bulk of current statistical research appears to be finding exact solutions to wrong problems instead of approximate solutions to right problems".

We will treat statistics in the very broad context of science and humanities and what were the past predictions in these areas and what predictions have come true. Perhaps, then we may be able to see our future role in time to come.

One thing is certain, that statistics has become essential in the interdisciplinary science, art and technology, that is at the heart of the scientific efforts in diverse fields, such as in industry,

commerce, government policy, making medical diagnosis,... Many academic disciplines also rely on statistics knowingly or unknowingly!

LASR Workshops started in 1973 to advance interdisciplinary research. We regard statistics as a continuously flowing river - with many currents, many directions, many waves. In our Workshop, we have crossed many rivers so far - from medical science to biological science, from image analysis to computer vision. (See the list of various LASR proceedings in the bibliography.) Very recent emphasis has moved to Bioinformatics and Wavelets.

In the vast ocean of knowledge, the journey and its joy will continue, we hope! In some sense, LASR Workshops are akin to the old Berkeley Symposiums headed by Jerzy Neyman, but the LASR Workshops are at a smaller scale. With changing technology, our Proceedings are available on web!

2 Theory/Models

Perhaps it will not be too far fetched to say that scientific research is driven by the search for theory to explain/predict a phenomenon. How statistics can help to show whether theory is true or not, i.e. a "hypothesis" is true or not. We are all aware of the Fisher paradigm that a test of significance when used accurately on a given data are capable of rejecting or "invalidating" hypotheses, but they are never capable of establishing them as certainly true. Thus a search for an alternative hypothesis becomes crucial. Rao (2001) describes the situation by an interesting conversation between Albert Einstein and some key statisticians.

Einstein: I have a new theory for explaining some natural phenomena. Can statisticians help in testing it?

Neyman and Pearson respond: Einstein, you have to do your own experiment, give us your data and also tell us what the possible alternatives are to your theory. We can tell you the most powerful method of verifying your theory.

Einstein: Alternative theories! There may be, but I do not know.

Fisher responds: I can give you the design of a perfect experiment to perform. The results can reject your theory if it is wrong and cannot confirm if it is true.

Wald and Wolfowitz respond: We would like to review your problem in terms of decision theory. Apart from other inputs, we need to know the losses involved in accepting and rejecting your theory.

Einstein: "If my theory is proven successful, Germany will claim me as a German and France will declare that I am a citizen of the world. Should my theory prove untrue France will say that I am a German, and Germany will declare that I am a Jew" (This is a true statement made by Einstein in an address at the Sorbonne).

Indeed, Einstein stated that

"no amount of experimentation can prove me right; a single experiment can prove me wrong."

Note that there is no conversation with De Finetti/D. Lindley but Rao (2001) claims that Bayesian methods cannot help here. "Bayesians argue that testing of hypothesis has no logical basis and that one should start with possible alternative hypotheses with known priori probabilities of being true and derive posterior probabilities in the light of the observed data. Since hypotheses not considered have zero prior probability, the true hypothesis when it is outside the chosen set of hypotheses (which generally happens) will never be discovered as it has zero posterior probability. "

We will come to various questions of statistical inference later on.

3 Explanation/Prediction

A related question to initial "theories/hypotheses" is of constructing the relevant model or models. In particular, statistical models which may or may not depend on underlying physical process.

Ripley (2003) has summarized the difference between explanation versus prediction as follows.

"For explanation, Occam's razor applies and we want

'an explanation that is as simple as possible, but no simpler' attrib Einstein

and we do have a concept of a 'true' model, or at least a model that is a good working approximation to the truth, for

'all models are false, but some are useful' G.E.P. Box.

Explanation is like doing scientific research.

On the other hand, prediction is like doing engineering development. All that matters is that it works. And if the aim is prediction, model choice should be based on the quality of the predictions."

In a broad context, there are many major problems needing suitable models. J.L. Casti in 1991 has posed the following challenging questions.

Weather and climate: Can we predict/explain changes in weather and climate?

Developmental biology: Can we predict/explain the development of the physical form of living things?

Stock markets: Can we predict/explain the behavior of stock market prices?

Warfare: Can we predict/explain the outbreak of war?

Mathematics: Can we predict/explain the true relations among numbers?

Some grades have been assigned by Casti (1991) along with a few areas given as benchmarks for comparison. These grades represent an ordinal representation of the degree to which science is able to capture the empirical evidence available for the phenomenon at hand. Economic theory scores very low but celestial machanics very high since we have now quite well- tested theories in this area. We still have to learn a lot about developmental biology.

4 Statistical Theology

In underlying statistical explanation/prediction there are three major competing philosophies:

Bayesian, frequentist and Fisherian.

These philosophies operate almost as theological schools of statistics, and a three-sided tug of war has been evident (more so in the past). Efron (1998) has given a barycentric picture showing the relative influence of these three philosophies upon various topics of modern statistical research. It is very much like a love triangle! Using what works is the best course. That is, be a pragmatic mathematical statistician!

D.R.Cox in Kardaun et al. (2003) has raised some deeper questions (14 in all) related to statistical inference. I am sure that these will keep statistical theorists busy for a long time to come!

1. How is overconditioning to be avoided?
2. How convincing is A.Birnbaum's argument that likelihood functions from different experiments that happen to be proportional should be treated identically (the so- called strong likelihood principle)?
3. What is the role of probability in formulating models for systems, such as economic time series, where even hypothetical repetition is hard to envisage?
4. Should nonparametric and semiparametric formulations be forced into a likelihood- based framework?
5. Is it fruitful to treat inference and decision analysis somewhat separately?
6. How possible and fruitful is it to treat quantitatively uncertainty not derived from statistical variability?
7. Are all sensible probabilities ultimately frequency based?
8. Was R.A Fisher right to deride axiomatic formulations (in statistics)?
9. How can the randomisation theory of experimental design and survey sampling best be accommodated within broader statistical theory?
10. Is the formulation of personalistic probability by De Finetti and Savage the wrong way round? It puts betting behaviour first and belief to be determined from that.
11. How useful is a personalistic theory as a base for public discussion?
12. In a Bayesian formulation should priors constructed retrospectively, after seeing the data, be treated distinctively?
13. Is the only sound justification of much current Bayesian work using rather flat priors the generation of (approximate) confidence limits? Or do the various forms of reference priors have some other viable justification?
14. What is the role in theory and in practice of upper and lower probabilities?

To me Question 6 is the key question. How statistics is related to fuzzy logic for example. In contrast, let me quote from an educational point of view,the answer by John Tukey to a question (Ferholz & Morgenthaler, 2003):

Q: The question at that time was - I was starting out in statistics - I asked you what I should read. You told me, the early Journal of the Royal Statistical Society, Supplement and Discussion. I found this a very important learning expreince and have also been using it in teaching. I was wondering if a young person came to you today, what would you tell them to start reading?
John Tukey: I guess you have to start at the same place, because to start anywhere else, you assume that they are a lot further along than you are when you start. And that one refers also to some of this thing about consulting, because the nearest thing to a surrogate for consulting that I know is to go and read the supplement to JRSS."

The supplement to JRSS has now become JRSS Series B.

5 Statistics and Computational Science

The influence of computers is now everywhere. In fact, statisticians were the first to use calculating machines from 1900. Karl Pearson had a large calculating laboratory to prepare various mathematical and statistical tables. In statistical practicals in universities, these calculating machines such as Facits were used even in the 1970's. Many interesting short cuts were known, e.g. to invert small matrices! (See for example, Mukhophadhyay, 2002).

At present, we are all influenced by Gordon Moor's Law (prediction) that computer hardware doubles its capability every eighteen months. As soon as you buy a computer, it is out of date!

In Mathematics, some theorems have been proved by computers (eg. the four-colour theorem) but because the proof involve many subcases (sometimes very complicated) that no person can check them! Some of the chess moves that Deep Blue made against Garry Kasparov had this same characteristic. Efron posed the following interesting scenario:

"Suppose that you could buy a really fast computer, one that could do not a billion calculations per second, not a trillion, but an infinite number. So after you unpacked it at home, you could numerically settle the Riemann hypothesis, the Goldbach conjecture, and Fermat's last theorem (this was a while ago), and still have time for breakfast. Would this be the end of mathematics?" - Efron (2002).

Fortunately, Hilbert in his 23 significant problems to solve did not mention any problem in statistical theory either! New challenges are outlined in Stewart (2002) as follows:

"In 2000 the Clay Mathematics Institute, in Cambridge, Massachusetts, offered prizes of $1 million each for solutions to seven long-standing and intractable mathematical problems. One is the Riemann hypothesis. The others are the Poincare conjecture, a topological characterization of the three-dimensional sphere; the P/NP problem of theoretical computer science, which asks for a proof that difficult computations really exist; the Hodge conjecture and the Birch/Swinnerton- Dyer conjecture in algebraic geometry; the existence (or not) of solutions to the Navier-Stokes equations of viscous fluid dynamics; and a proof of the "mass gap hypothesis" in quantum field theory."

There are really no equivalent long-standing challenges in Statistics. However, there may be a chance to solve these by computers. Perhaps more so when computers develop "common sense" (or intuition). Minsky (2003) muses on this aspect as follows:

"The main thing that has remained the same is that computers still know so little about their world. In particular, they have no idea about the goals of the people who use them. This is why, for example, most programs will die whenever their users make a mistake, be it a grave conceptual error or just typing an incorrect character. Someday, however, computers will have the sorts of common-sense knowledge that most of us share - millions of everyday facts about the world and common-sense ways to think about them. If they learn to think about themselves and invent new ways to improve themselves, then everything we know will change and (if we can keep control of them) we'll never need to work again."

6 Some recent Statistical Achievements

The development of statistics in the early 20th century was driven by a small number of areas (agriculture, astronomy, official statistics, elementary biology) but subsequently statistics has become a central tool for many areas. We have witnessed in our LASR workshop the emergence of new areas through interdisciplinary research, areas such as spatial statistics, statistical

simulation, large scale data as in imaging, machine learning, neural nets, MCMC, financial derivatives, bioinformatics, shape analysis and so on.

Green (2003) has nicely summarized on the present omnipresence of statistics.

"As we all recognize, statistics is an extraordinary discipline. It reaches out into business and industry, into public life and into most other disciplines. In turn, these interactions have profoundly shaped the subject. It embraces a huge variety of aspects: philosophical foundations, mathematical theory, inferential principles, design, data collection, techniques, computation, modelling, and so on. And crucially, the conduct of its interactions with the rest of the world is part of the subject itself."

However there is definitely image problem. Again Green (2003) has described it succinctly.

"Some users of statistical methods from outside the discipline apparently regard statistics as a static thing, a shelf-full of technique and good practice gathering dust, to be consulted only occasionally, and without enthusiasm or much engagement. In a world in which 'information' has become both a global currency and a global product, in which 'uncertainty' is undiminished and its impact more widely appreciated and in which 'quantification' rules, it is astonishing that a discipline whose centre-piece is the quantification of uncertain information should have this image."

7 A long term forecast

Here we start with a prediction by Arthur Clark who gave a list of what are the past achievements together with his prediction in 1960. Some of these have been already achieved such as bio-engineering predictions and planetary landings. Some are still to be achieved such as suspended animation and immortality!

We again start with prediction in science. This time by Kurzweil (1999). According to Kurzweil, by 2099, we will have no clear distinction between humans and computers! What a prospect!!

I personally believe that understanding 'consciousness' is one of the major challenges. Let me quote Holland (2002);-

"Attempts to discover mechanisms that generate thought and consciousness have occupied humankind since the beginning of recorded history. Most psychologists now believe that consciousness is tied to the activity of neurons in the central nervous system, but we still know surprisingly little about the relation between consciousness and neural activity. Unraveling this relation has proved to be notoriously difficult, and I do not expect sudden "solutions" in the next fifty years."

However, Crick (1995) postulates what he calls The Astonishing Hypothesis:-

"The Astonishing Hypothesis is that 'You', your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more that the behaviour of a vast assembly of nerve cells and their associated molecules. As Lewis Carroll's Alice might have phrased it: 'You're nothing but a pack of neurons.' This hypothesis is so alien to the ideas of most people alive today that it can truly be called astonishing."

I do not see any statistics directly there but it is hidden through neural nets, and statisticians have to be aware of these future directions!

We now move to some other intellectual pursuits. Here Richard Feynman's (in Strotgatz, 2002) challenge is worth repeating:

"The next great era of awakening of human intellect may well produce a method of under-

standing the qualitative content of equations. Today we cannot. Today we cannot see that the water flow equations contain such things as the barberpole structure of turbulence that one sees between rotating cylinders. Today we cannot see whether Schrodinger's equation contains frogs, musical composers, or morality - or whether it does not."

Strogatz (2002) adds:
"If we're ever going to reach that next great era of awakening, we'll need to be rescued from the devil of dimensionality. Look for computers to be our saviours. Once they become reasonably intelligent, they should be able to visualise any number of dimensions. They already do the grunt work of running our simulations; maybe the day will come when they will also extract the laws of self- organisation in complex systems."

We are quite familiar with this curse of dimensionality in multivariate analysis.

I believe that quantitative biology has many challenges(see, in particular, Gilks, 2004). In data mining, statistical revolution has not yet happened. Topics such as non-Euclidean statistics (including shape analysis), False discovery rate, Six sigma technology (aiming for not more than 3.4 defects in 1 million, see Hahn et al 1999) will develop further. Modular statistics (graphical models/hierarchical models) is another large area. There is a strong similarity of Modular Statistics with Complex Adaptive Systems (CAS) where computer scientists have made a headway.

"Complex adaptive systems (CAS) consist of many interacting components, called agents, that adapt to (or learn from) each other as they interact. Stock markets and immune systems are familiar examples of CAS. Even on relatively short timescales, CAS exhibit a range of nonadditive (nonlinear) effects: self-organisation, chaos, fractal attractors, frozen accidents, level points, and the like. As a result, the actions of the components cannot be summed to give an overall trend. Moreover, we have only bits and pieces of a theory of CAS. Because we lack an overreaching theory, we have no general, principled way for determining the influence of these nonadditive effects. As a consequence, when CAS are involved, prediction is fraught with hazard."

However there are many challenges in CAS: The first challenge is to characterize the connectivity of complex networks something beyond the small-world phenomenon (producing six degrees of separation). We need to learn how real networks are actually structured to understand how the brain computes or why cells become cancerous. The second challenge is of finding the rules governing the interactions between genes, people, or companies. To quote Holland (2002):

"Our models of complex systems will never advance beyond caricatures until we can find a way to infer local dynamics from data. For instance, consider the genetic networks that control the workings of a cell. With the DNA chips now available, we can measure the simultaneous activity of thousands of different genes as a function of time, but we still don't know which genes are talking to which and how they are influencing each other's activity in a quantitative way. .... If we can develop systematic ways to infer dynamics from Multiple Time- series Measurements, that advance will have tremendous implications - not just for biology but for sociology and economics as well."

Thus spatial temporal modelling through graphical models must be a rich area!

Raftery et al (2003, Introduction) have taken a broader view of the future of statistics:-

"So where is the field of statistics going in the new millennium? While prediction is hard, especially about the future, it does seem safe to say that new developments will be driven by new kinds of data requiring analysis and by the development of computing to make them possible. Gene expression data is one current example of this, and this is a field where statisticians have rapidly become deeply involved. Datamining is another; this started life as the analysis of

retail barcode data, and statisticians have become involved more slowly there. One area where statistics has been largely absent, but where new theory and computing power may allow it to make a contribution, is the analysis of simulation or mechanistic models, which are mostly deterministic and dominate scientific endeavour in many disciplines, often largely to the exclusion of more conventional statistical models."

A statistical study of mechanistic models or "Predictive Statistics" must be another key area of future development. Our Statistical research should be coupled with the following change of attitude as outlined by Green (2003) in his presidential address to the Royal Statistical Society:

"If we accept the idea that the dissemination of statistical methodology research has at least two components - the paper and the reference implementation - then it is a logical next step to adopt a good idea from current practice in the life sciences, that of the Web page supporting the paper, whereon the published paper is available, together typically with a more discursive version of the work, giving more detail than space allows in a journal, and access to the data sets used in illuminations. This is a convenient place to put the reference code and instructions for users."

Fisher (1953) made the following optimistic statement in his presidential address to the Royal Statistical Society:-

"I venture to suggest that statistical science is the peculiar aspect of human progress which gives to the twentieth century its special character; and indeed members of my present audience will know from their own personal and professional experience that it is to the statistician that the present age turns for what is most essential in all its more important activities."

Hopefully, it will apply to many more centuries! We do not want future statisticians to be out of a job!!

References

Aykroyd, R.G., Mardia, K.V. and Langdon, M.J. (2003) eds *Stochastic Geometry, Biological Structure and Images*. LASR Proceedings, Leeds Univ. Press, Leeds.

Aykroyd, R.G., Mardia, K.V. and McDonell, P. (2002) *Statistics of Large Data Sets: Functional and Image Data, Bioinformatics and Data Mining*. LASR Proceedings, Leeds Univ. Press, Leeds.

Brockman, J. (ed) (2003).*The New Humanists - Science at the Edge*. Barnes and Noble, New York.

Casti, J.L. (1991) *Searching for Certainty*. Abacus, London.

Crick, F. (1994) *The Astonishing Hypothesis*. Simon and Schuster Ltd.

Efron, B. (1998) R.A. Fisher in the 21st Century(with discussion). *Statistical Science*, **13**, pp.95-122.

Efron, B. (2002) The bootstrap and modern Statistics in *Raftery et al*. (2002) below. pp.326-332.

Fisher, R.A. (1953) The expansion of statistics. *J. Roy. Statist. Soc.* A, **116**, pp.1-6.

Ferholz, L.T. and Morgenthaler, S. (2003) A conversation with John W. Tukey. *Statistical Science*, **18**, pp.346-356 (Reprint of 1997 in the Practice of Data Analysis).

Gilks, W. (2004). Bioinformatics: new science- new statistics. *Significance*, **1**, pp.7-9.

Green, P.J. (2003) Diversities of gifts, but the same spirit. *The Statistician*, **52**, pp.423-438.

Hahn, G.J., Hill, W.J., Hoerl, R.W. and Zinkgraf, S.A. (1999) The impact of six sigma improvement - A glimpse into the future of statistics. *Amer. Statistician*, **53**, pp.208-215.

Holland, J.H. (2002) *What is to come and how to predict it*. In *Brockman* (2002) above, pp. 170-182.

Kardaun, O.J.W.F., Salome, D., Schaafsma, W., Steerneman, A.G.M., Willems, J.C. and Cox, D.R. (2003) Reflections on fourteen cryptic issues concerning the nature of statistical inference. *International Statistical Review*, **71**, pp.277-318.

Kent, J.T. and Aykroyd, R.G. (2000) (eds) *The Statistics of Directions, Shapes and Images*. LASR Proceedings, Leeds University.

Kurzweil, R. (1999) *The Age of Spiritual Machines*. Orion Business Books, London.

Mardia, K.V. and Aykroyd, R.G. (2001) (eds) *Functional and Spatial Data Analysis*. LASR Proceedings. Leeds Univ. Press, Leeds.

Mardia,K.V., Aykroyd,R.G. and Dryden,I.L. (1999) (eds) *Spatial Temporal Modelling and its Applications*. LASR Proceedings, Leeds Univ. Press, Leeds.

Mardia, K.V., Gill, C.A. and Aykroyd, R.G. (1997) (ed.) *The Art and Science of Bayesian Image Analysis*. LASR Proceedings. Leeds Univ. Press, Leeds.

Mardia, K.V., Gill, C.A. and Dryden, I.L. (1996) (eds.) *Image fusion and shape variability techniques*, LASR Proceedings. Leeds Univ. Press, Leeds.

Mardia, K.V. and Gill, C.A. (1995) (eds.) *Current Issues in Statistical Shape Analysis*, LASR Proceedings, Leeds Univ. Press, Leeds.

Minsky, M. (2003) What comes after minds? In *Brockman* (2003) above. pp.197-214.

Mukhopodhyay.N.(2002) A conversation with Kanti Mardia. *Statistical Science*, **17**, pp.113-148.

Raftery, A.E., Tanner, A.E. and Wells, M.T. (2001) *Statistics in the 21st Century*. Chapman and Hall, London.

Rao, C.R. (2001) Statistics: Reflections on the past visions of the future. *Com. Statist. Theory Methods*, **30**, pp.2235-2257.

Rao, C.R. and Szekely, G.J. (eds), (2000), *Statistics for the 21st Century* . Marcel Delcker. New York.

Ripley, B.D. (2003) Model selection in complex classes of models. "Statistical learning". AMSI meeting at UNSW. http://www.stats.ox/∼ripley/talks.html.

Stewart, I. (2002) The Mathematics of 2050. In *Brockman* (2002) above pp.29-40.

Strogatz, S. (2002) "Fermi's Little Discovery" and the future of chaos and complexity theory. In *Brockman* (2002) above pp. 114-125.